Training and spontaneous reinforcement of neuronal assemblies by spike timing

Gabriel Koch Ocker[1,3,4], Brent Doiron[2,3]

1: Department of Neuroscience, University of Pittsburgh, Pittsburgh, PA, USA

2: Department of Mathematics, University of Pittsburgh, Pittsburgh, PA, USA

3: Center for the Neural Basis of Cognition, University of Pittsburgh and Carnegie Mellon University, Pittsburgh, PA, USA

4: Allen Institute for Brain Science, Seattle, WA, USA

1 Abstract

The synaptic connectivity of cortex is plastic, with experience shaping the ongoing interactions between neurons. Theoretical studies of spike timing–dependent plasticity (STDP) have focused on either just pairs of neurons or large-scale simulations where analytic insight is lacking. A simple account for how fast spike time correlations affect both micro- and macroscopic network structure remains lacking. We develop a low-dimensional mean field theory showing how STDP gives rise to strongly coupled assemblies of neurons with shared stimulus preferences, with the connectivity actively reinforced by spike train correlations during spontaneous dynamics. Furthermore, the stimulus coding by cell assemblies is actively maintained by these internally generated spiking correlations, suggesting a new role for noise correlations in neural coding. Assembly formation has been often associated with firing rate-based plasticity schemes; our theory provides an alternative and complementary framework, where temporal correlations and STDP form and actively maintain learned structure in cortical networks.

2 Introduction

A cornerstone principle that bridges systems and cellular neuroscience is that the synaptic wiring between neurons is sculpted by experience. The early origins of this idea are often attributed to Donald Hebb [1–3], who famously postulated that groups of neurons that are repeatedly coactivated will strengthen the synaptic wiring between one another. The interconnected group, termed an *assembly*, has become an essential building block of many theories of neural computation [4] and associative memory [5]. Despite the functional appeal of neuronal assemblies, only recently has physiological evidence of assembly structure been collected.

In mouse primary visual cortex, new advances in recording techniques have shown that pyramidal neurons with similar stimulus preferences connect more frequently, with more synapses and with stronger postsynaptic potentials than neurons with dissimilar stimulus preferences [6–8]. Synaptically connected neurons tend to receive more common inputs than would be expected by chance, suggesting a clustered architecture [9, 10]. While strong recurrent connectivity between

similarly tuned neurons is present even at eye opening, it is enhanced during development and especially by visual experience [11, 12]. This suggests long-term synaptic plasticity as a key mechanism for the assembly organization of cortical circuits. However, we have only a partial understanding about how the mechanics of synaptic plasticity interacts with recurrent circuits to support the training and maintenance of assembly structure.

Physiological investigation over the past two decades has uncovered spike timing–dependent plasticity (STDP) mechanisms whereby the temporal correlations of pre- and postsynaptic spiking activity drive learning [13]. Hebbian STDP reinforces temporally causal interactions between neurons: the connections from presynaptic neurons that causally contribute to a postsynaptic neuron's firing are strengthened, while the other connections are weakened. Consequently, many modeling studies show that Hebbian STDP promotes the development of feedforward networks [14–16] with temporally precise [17] and tuned [18] responses giving rise to sequential activity [19, 20]. Feedforward structures are quite distinct from the recurrent wiring within neuronal assemblies and thus it is not obvious that STDP will support assembly formation. Nevertheless, recent theoretical work has shown that networks of recurrently coupled spiking neurons having STDP in excitatory connections effectively learn assembly structure [21–23] that is stable in the face of ongoing spontaneous spiking activity post training [22, 23].

The synaptic plasticity models [24–26] used in these studies [22, 23] capture the known firing rate dependence of the balance between potentiation and depression [27]. When spike time correlations are neglected, these models admit reductions of STDP learning to more classic rate-based plasticity schemes [24, 28, 29] so that when high (low) postsynaptic activity is paired with high presynaptic activity, synaptic connections are potentiated (depressed) (Figure 1A). In these models, the assembly structure in recurrent networks forms via firing rate transitions that toggle between strongly potentiation- and depression-dominated regimes (Figure 1B). While these past studies [22, 23] show that assembly formation can co-occur with STDP, in these networks any fast spike time correlations between neurons contribute minimally to synaptic learning.

Spike trains in diverse cortical areas do exhibit covariable trial-by-trial fluctuations (noise correlations). These noise correlations covary with neurons' stimulus preferences (signal correlations) [30–32] and synaptically connected neurons have higher noise correlations [6, 7], suggesting that assembly structure and noise correlations are related. Furthermore, excitatory-inhibitory interactions within cortical circuits create nearly synchronous (~ 10 ms) joint temporal structure between spike trains that overlaps with the fine timescale required for STDP learning [33–35]. In complementary research, several in vivo studies show that the precise timing of pre- and postsynaptic spikes can be a crucial determinant of plasticity [36]. In particular, Kim et al. [37] have recently shown that correlations on the order of tens of milliseconds control assembly formation in vivo. Thus, while previous modeling studies did not require fast spike train correlations to

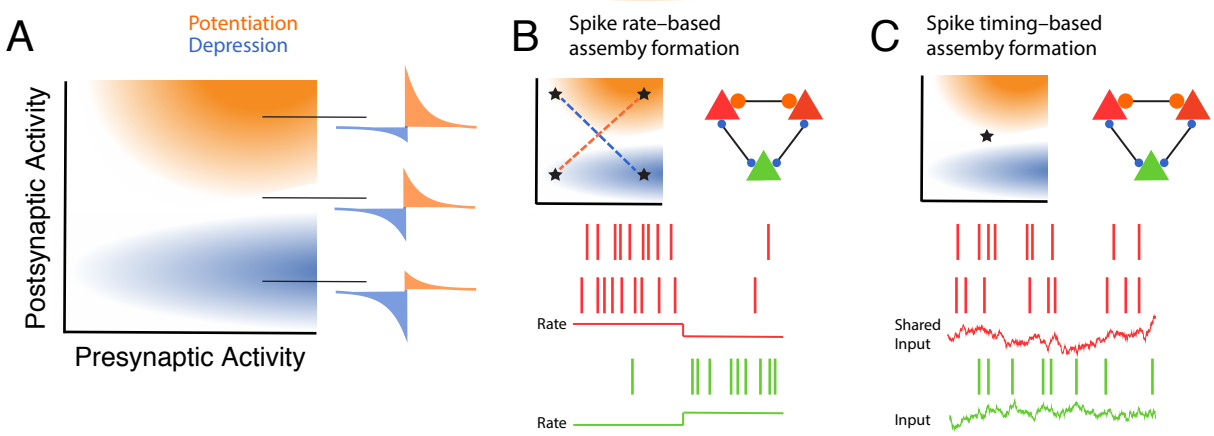

Figure 1. Spike rate– versus spike timing–based neuronal assembly formation. (A). Schematic illustrating how combinations of pre- and postsynaptic activity combine to drive synaptic potentiation and depression in models of STDP. The schematic is adapted from Litwin-Kumar & Doiron [22] where STDP rules based on third-order spike interactions [24], or voltage- [29] or calcium-based [26] learning were studied. Pre- and postsynaptic correlations are neglected. The STDP curves on the right indicate the degree of potentiation and depression as pre- and postsynaptic activity ranges. (B) Example three neuron group with two neurons having co-fluctuating firing rates (bottom, red–red) and the other neuron having anti-correlated firing rate fluctuations (bottom, green–red). This dynamic potentiates synaptic coupling between correlated neurons while depressing synaptic coupling between anti-correlated neurons (right graph). (C) Same as B except firing rates are fixed at a value that balances rate-based potentiation and depression. Shared input correlations to two neurons can potentiate strong recurrent synapses (bottom, red-red) and depress uncorrelated neurons (bottom, green-red).

train stable assembly structure [22, 23], there is sufficient experimental evidence to suggest that spike timing may nonetheless play an important role in assembly formation and stability.

Here, we show that spike time correlations can, in the absence of rate-based plasticity mechanisms, form Hebbian assemblies in response to correlated external inputs to groups of neurons (Figure 1C). We extend past studies [16, 28, 38] and combine linear response theory for spike train covariances in recurrent networks [39] with a slow-fast theory of STDP [40] to develop low-dimensional theories describing the evolution of the network structure. Our analyses reveal that training promotes strong connectivity and strong *reciprocal* connectivity within co-stimulated groups. We further show that after training and in the absence of any external input correlations, internally generated spike time correlations reinforce learned architectures during spontaneous activity. Finally, this result motivates us to speculate on a new beneficial role of internally generated noise correlations on stimulus coding: to maintain stimulus-specific assembly wiring that supports enhanced response sensitivity. In total, our theory reveals a potential role for precise spike time correlations in the formation of neuronal assemblies in response to correlated external inputs, as well as their active maintenance during spontaneous activity.

3 Results

3.1 Plasticity of partially symmetric networks during spontaneous activity

We first present the basic network properties of our network (see Methods 5.1). One striking feature of cortical networks is the overrepresentation of reciprocally connected pairs of excitatory neurons, compared to a simple randomly wired (Erdős-Rényi) network [10, 41]. In order to reflect this structure, we took the baseline excitatory-excitatory connectivity of our network, $\mathbf{W}^0_{EE}$, to be composed of two parts: $\mathbf{W}^0_{EE} = \mathbf{W}^0_{\text{sym}} + \mathbf{W}^0_{\text{asym}}$, where $\mathbf{W}^0_{\text{sym}}$ is a symmetric random binary matrix with connection probability Ωp_0 and $\mathbf{W}^0_{\text{asym}}$ a random binary matrix with connection probability $(1-\Omega)p_0$ (without any symmetry constraint). Both had Erdős-Rényi statistics. The parameter Ω thus determined the frequency of bidirectionally connected pairs of excitatory neurons in $\mathbf{W}^0_{EE}$. We modeled networks of 1500 excitatory neurons and 300 inhibitory neurons, both types following exponential integrate-and-fire dynamics [42]. The overall connection probability between excitatory neurons was $p_0 = 0.15$, with $\Omega = 0.4$. Excitatory-inhibitory, inhibitory-excitatory and inhibitory-inhibitory connectivity were asymmetric ($\Omega = 0$), with connection probability 0.4.

Before we proceed to the formation of assembly structure, we present the underlying synaptic dynamics of both the excitatory-excitatory and inhibitory-excitatory connections in the network in the absence of a training signal.

3.1.1 Excitatory plasticity and thresholds for synaptic weight dynamics

In this study we consider the evolution of the weighted connectivity $\mathbf{W}_{EE}$ (Figure 2A). In order to focus on learning due to precise spike time correlations, we used a classical Hebbian spike pair–based plasticity rule for the plasticity between excitatory neurons (eSTDP) [17, 43, 44] (Figure 2B). The plasticity rule is phenomenological, and embodies the simple observation that spike pairs induce changes in synaptic weights and the amplitude of these changes depends on the time lag between the two spikes [13]. The coupling strength scaled with system size as $\epsilon = (Np_0)^{-1}$ so that for large N the evolution of $\mathbf{W}_{EE}$ was slow compared to the fast timescales of membrane dynamics and spike discharge (Figure 2C). The separation of timescales between spike time and synaptic weight dynamics permitted an averaging theory for the joint dynamics of $\mathbf{W}_{EE}$ and the spike time covariance $\mathbf{C}(s)$ (see [38] for a full description).

We began with a simple characterization of the network excitatory-excitatory structure in terms of two variables:

$$\begin{aligned} \epsilon p &= \frac{1}{N_E^2} \sum_{i,j \in E} \mathbf{W}_{ij} \\ \epsilon q &= \frac{1}{N_E^2} \sum_{i,j \in E} \mathbf{W}^0_{ij} \mathbf{W}_{ji} - \epsilon p_0 p \end{aligned} \tag{1}$$

These measure the mean weight of excitatory-excitatory synapses (p) and the mean weight of *reciprocal* excitatory-excitatory synapses (q) above what would be expected in an unstructured network. (Here, q corresponds to $q_{\mathrm{X}}^{\mathrm{rec}}$ in [38]). Note that with asymmetric connectivity, $\Omega = 0$, q becomes weak ($\mathcal{O}(N^{-3/2})$) so that the network connectivity can be described (to leading order) only by p. The structure we impose on the network by setting $\Omega \neq 0$ enforces that the variables p, q form, to leading order, an invariant set for the plasticity of synaptic motifs [38].

We derived dynamics for these variables following the same steps as in [38] (see Methods, 5.3). We first approximated the average spike train covariance from the contributions of length one paths in the network and neglected the bounds on synaptic weights in the eSTDP rule, so that this theory does not account for equilibrium states of the weights. The network structure p, q then obeys:

$$\frac{dp}{dt} = \left(r_E^2 S + c_{EE}\sigma^2 S_\eta\right) p_0 + \epsilon \left[S_F p + S_B \left(q + p_0 p\right) + S_C p^2 + S_C^I \gamma (p_{EI}^*)^2\right] \tag{2}$$

$$\frac{dq}{dt} = \left(r_E^2 S + c_{EE}\sigma^2 S_\eta\right) q_0 + \epsilon \left[S_F q + S_B \left(1 - p_0\right)\left(q + p_0 p\right) + S_C \frac{q_0}{p_0} p^2 + S_C^I \gamma \frac{q_0}{p_0} (p_{EI}^*)^2\right] \tag{3}$$

The first terms on the right-hand side of Eq. (2) describe the contributions of chance spike coincidences ($r_E^2 S$), with r_E being the network-averaged firing rate, and correlations induced by external inputs ($c_{EE}\sigma^2 S_\eta$). S is the integral of the eSTDP rule, while S_η is the integral of the eSTDP rule against the average susceptibility of two neurons to externally induced correlations (Methods, 5.3). The latter terms describe the contribution of correlations induced by coupling within the network, weighted by the eSTDP rule. The effect of correlations due to direct (forward) connections is measured by S_F, and those due to reciprocal (backward) connections is measured by S_B. The final terms arise from correlations due to common inputs from excitatory (S_C) or inhibitory (S_C^I) neurons. The parameter γ is the ratio of the number of inhibitory neurons to excitatory neurons (here $\gamma = 1/3$) and we defer a treatment of the inhibitory to excitatory connection strength p_{EI}^* until the next section. Finally, q_0 is the empirical frequency of reciprocal synapses in the network above chance levels, analogous to q but measured from the adjacency matrix rather than the weight matrix.

We took there to be a balance between potentiation and depression, so that $S \sim \mathcal{O}(\epsilon)$ (star in Figure 1C), with that balance tilted slightly in favor of depression (so that $S < 0$). This assumption, when combined with an absence of training ($c_{EE} = 0$), leads to the synaptic dynamics being governed by different sources of internally generated spiking covariability, each interacting with the eSTDP rule $L(s)$. Spiking covariations from direct connections mainly contribute at positive time lags, interacting with the potentiation side of the eSTDP rule. This is reflected in the average spike train covariance between monosynaptically connected neurons (Figure 2F, left). Reciprocal connections, in contrast, contribute spiking covariations at negative

time lags, interacting with the depression side of the eSTDP rule. This is reflected in the average spike train covariance between reciprocally connected pairs, which includes the contributions from both direct and reciprocal connections (Figure 2F, middle). Finally, the contributions from common inputs are temporally symmetric around zero time lag, interacting with both the potentiation and depression windows. The average spike train covariance between all neurons was asymmetric because of the higher frequency of monosynaptically connected over reciprocally connected neurons (Figure 2F, right).

The competition between these sources of spiking covariability imposed thresholds for potentiation and depression of the mean field variables p and q. This is best understood by examining the p and q nullclines (Figure 2D, blue lines). Briefly, a nullcline is the collection of (p, q) values where (for example) $dp/dt = 0$; on either side of the nullcline the sign of dp/dt dictates the evolution for p. The nullclines of Eqs. (2) and (3) intersect at a single point in (p, q) space, and for the Hebbian plasticity rule used (Figure 2B) that point was an unstable repeller, with dynamics flowing away from the point (Figure 2D, red arrows). In this case the nullclines then acted as thresholds so that if either p or q were initially stronger than its threshold it would potentiate, and otherwise it would depress. It has been long known that additive Hebbian eSTDP produces unstable synaptic dynamics for pairs of coupled neurons through a competition between potentiation and depression [45, 46]. Our theory has extended this idea to large populations of neurons through mean field dynamics of p and q.

Our theory not only gives a qualitative understanding of synaptic dynamics, it also provides a good quantitative prediction of the plasticity within our large-scale integrate-and-fire network (Figure 2E, compare the solid theory curves to the shaded curves estimated from numerical simulations). The threshold dynamics for p and q, and their dependence upon various aspects of spike time correlations, will serve as an important component of assembly formation. Before examining how external input correlations can train the network into different macroscopic structures, we first must examine the role of inhibition and inhibitory plasticity in this network.

3.1.2 Inhibition and homeostatic inhibitory STDP maintain stable activity

In recurrent networks, excitatory plasticity can lead to the destabilization of asynchronous activity [47] and the development of pathological synchrony [48]. Past modeling studies have explored plasticity of inhibition as a stabilizing mechanism [49], preventing runaway activity in networks with [22, 23] and without [50] excitatory plasticity. Recent experiments in humans using a combination of transcranial direct current stimulation and ultra-high field MRI has given evidence for an association-dependent balancing of excitation and inhibition [51], where inhibitory plasticity was a suggested mechanism. Indeed, plasticity of inhibitory-excitatory connectivity maintains a balance between excitation and inhibition in layer 5 of mouse auditory cortex in vitro [52]. We followed these studies and, to prevent runaway excitation due to potentiation of

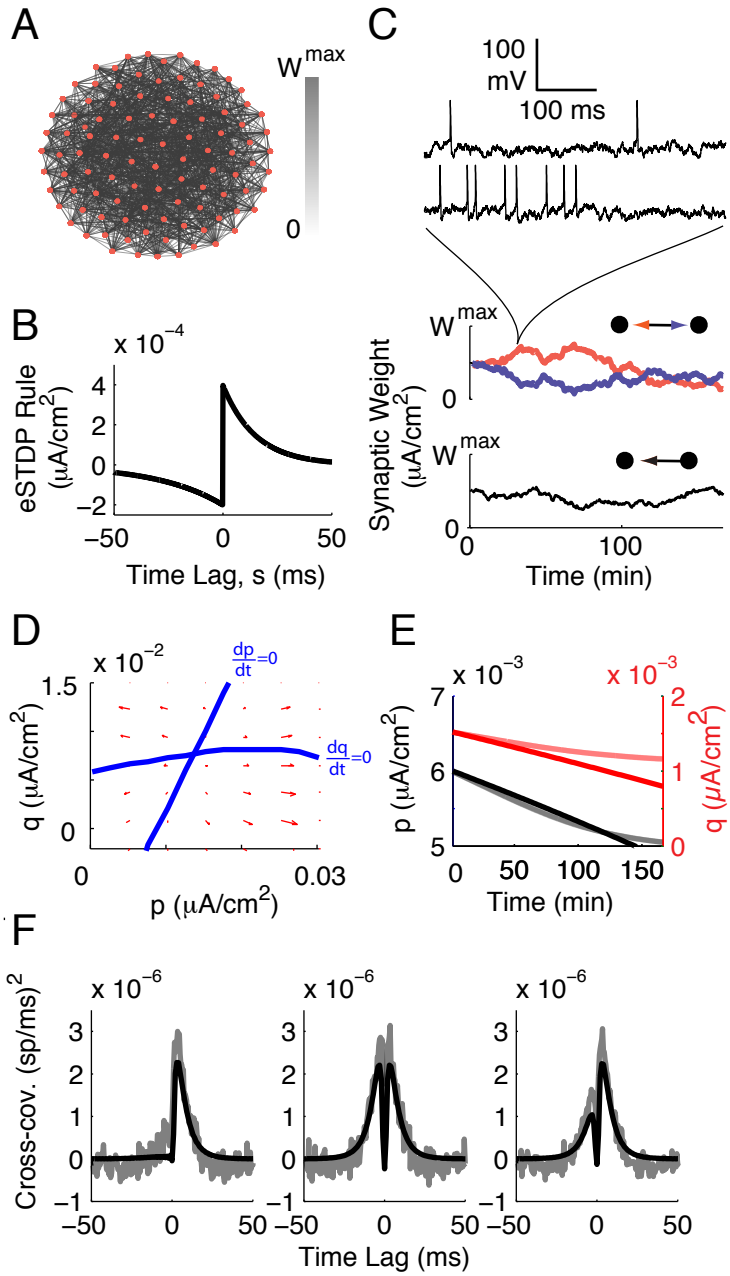

Figure 2. Network structure shapes synaptic plasticity. (A) Visualization of a random subset of the excitatory-excitatory connectivity. (B) The eSTDP rule, $L(s)$, is composed of exponential windows for depression (-) and potentiation (+). Each is defined by its amplitude $f_\pm$ and timescale $\tau_\pm$. (C) Synaptic weights evolve on a slow timescale. Individual synaptic weights are governed by the relative timing of spikes in the pre- and postsynaptic neurons' spike trains. (D) Dynamics of the mean synaptic weight (p) and the mean above-chance strength of *reciprocal* synapses, q. There is a threshold for potentiation of each given by its nullcline (blue lines). (E) Time course of p and q in the case where both are depressing. Solid lines: theory, Eqs. (2),(3). Shaded lines: simulation of the spiking network. (F) Average spike train covariance between monosynaptically connected pairs (left), reciprocally connected pairs (right) and all pairs (right). Shaded lines: simulation. Solid lines: linear response theory (first-order truncation, Eq. (23)).

excitatory synapses, we modeled inhibitory → excitatory homeostatic spike timing–dependent plasticity (iSTDP): pairs of near coincident pre- and postsynaptic spikes caused potentiation of inhibitory-excitatory synapses, while individual presynaptic spikes caused depression [50] (Figure 3A). The strength of this depression was determined by the homeostatic target excitatory rate, $\bar{r}_E$ (Methods, 5.2).

We took the excitatory eSTDP rule to be balanced between potentiation and depression, but if the excitatory firing rates were far from the target rate $\bar{r}_E$, then the inhibitory plasticity became unbalanced and its leading-order dynamics did not depend on internally generated spike time correlations (Methods, 5.5):

$$\frac{dp_{EI}}{dt} = \left(r_I\big(r_E - \bar{r}_E\big)S^I + c_{EI}\sigma^2 S_\eta^{EI}\right)p_0^{EI} \tag{4}$$

Together with the dynamics of the firing rates r_E, r_I, these occurred on a faster timescale than the balanced plasticity of excitatory connectivity. Examining the fixed points and stability of (p_{EI}, r_E, r_I) on this unbalanced timescale revealed that the inhibitory plasticity stabilizes the firing rates so that $r_E - \bar{r}_E \sim \mathcal{O}(\epsilon)$ (Methods, 5.5). Indeed, in simulations we saw that as p increased (decreased), p_{EI} potentiated (depressed) and maintained $r_E = \bar{r}_E + \mathcal{O}(\epsilon)$ (e.g., Figure 3B).

The location of the homeostatic inhibitory weight, p^*_{EI}, is given by solving the leading-order dynamics of the unbalanced inhibitory plasticity for $dp_{EI}/dt = 0, dr_E/dt = 0, dr_I/dt = 0$. Due to the separation of timescales between the homeostatic iSTDP and the balanced eSTDP, we could predict the location of the homeostatic inhibitory weight p^*_{EI} through a quasi-static approximation of p (Methods, 5.5). We tracked the location of the homeostatic inhibitory weight p^*_{EI} as a function of p. As expected, strong recurrent excitation required stronger inhibitory-excitatory feedback to enforce $r_E = \bar{r}_E + \mathcal{O}(\epsilon)$ (Figure 3C). In order to investigate the conditions under which inhibition was able to maintain stable activity at that homeostatic fixed point, we compared the cases of plastic and nonplastic inhibition. With nonplastic inhibition, firing rates increased with p. If the excitatory feedback p became strong enough, the stationary firing rates lost stability (Figure 3D). This instability was reflected in the development of hypersynchronous spiking (Figure 3E), in contrast to the weakly correlated spiking activity in the network with plastic inhibition. In total, in order to study the robust potentiation of recurrent excitation, we required a counterbalancing potentiation of inhibitory onto excitatory neurons so as to homeostatically maintain a weakly correlated yet strongly connected excitatory network.

3.2 Stimulus-induced correlations drive assembly formation

The thresholds for potentiation and depression in both p and q suggested a mechanism for the formation of assembly structure through spike timing. Namely, if we define p and q variables

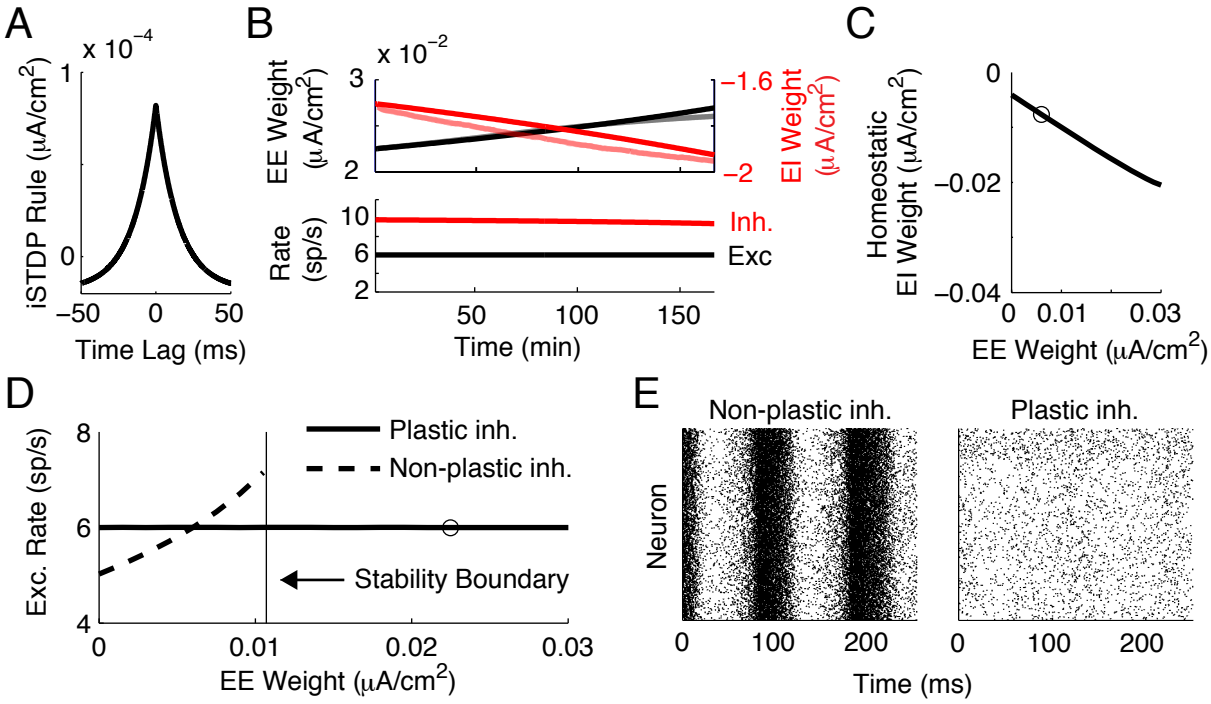

Figure 3. Homeostatic inhibitory plasticity dynamically stabilizes firing rates. (A) STDP rule for inhibitory-excitatory synapses. (B) Top: Coevolution of mean excitatory-excitatory weight p (black) and mean inhibitory-excitatory p_{EI} (red). Bottom: Firing rates during plasticity. (C) The fixed point for p_{EI} as a function of the mean excitatory strength p. Open circle marks the inhibitory weight used for the nonplastic inhibition in later panels. (D) Firing rates as a function of excitatory weight in the cases of plastic and nonplastic inhibition. We predicted the location of that stability boundary by numerically computing the eigenvalues of the Fokker-Planck equation associated with the single-neuron voltage distribution and examining how activity is recurrently filtered through the network [53]. (E) Raster plots of the network activity. In both bases the excitatory weight is at the value marked by the circle in panel D. For the right raster, p_{EI} is at its homeostatic fixed point.

for within- and cross-assembly connectivity, each should obey similar dynamics to Eqs. (2), (3). In particular, each should have a threshold for potentiation. Furthermore, these thresholds should depend on the spatial correlation of the external inputs to within- or cross-cluster pairs of neurons.

We began by studying the simpler case of networks with asymmetric baseline connectivity ($\Omega = 0$) so that q could be neglected. We divided the excitatory neurons into M putative assemblies of κ neurons each, based on their assigned stimulus preferences. Each assembly contained neurons that received spatially correlated inputs due to an external stimulus (Figure 4A). For ease of calculation, we assumed that the assemblies were symmetric so that the

connectivity within and between assemblies was characterized by:

$$\begin{aligned} \epsilon p_{AA} &= \frac{1}{\kappa^2} \sum_{i,j \in A} \mathbf{W}_{ij} \\ \epsilon p_{AB} &= \frac{1}{\kappa(N_E - \kappa)} \sum_{i \in A} \sum_{j \notin A} \mathbf{W}_{ij} \end{aligned} \tag{5}$$

where p_{AA} is the mean strength of connections within an assembly, and p_{AB} is the mean strength of all cross-assembly connections. The correlation of the external inputs to neurons with the same (different) input preferences was c_{AA} (c_{AB}). The inhibitory-excitatory, excitatory-inhibitory and inhibitory-inhibitory connectivities remained unstructured and asymmetric. Following the same steps as for p, we derived dynamical equations for the mean within- and cross-cluster connectivity (Methods, 5.6):

$$\frac{dp_{AA}}{dt} = \left(r_E^2 S + c_{AA}\sigma^2 S_\eta\right) p_0 + \epsilon\left[S_F p_{AA} + S_B p_0 p_{AA} + S_C\left(p_{AA}^2 + (M-1)\, p_{AB}^2\right) + S_C^I \gamma (p_{EI}^*)^2\right] \tag{6}$$

$$\frac{dp_{AB}}{dt} = \left(r_E^2 S + c_{AB}\sigma^2 S_\eta\right) p_0 + \epsilon\left[S_F p_{AB} + S_B p_0 p_{AB} + S_C\left(2 p_{AA} p_{AB} + (M-2)\, p_{AB}^2\right) + S_C^I \gamma (p_{EI}^*)^2\right] \tag{7}$$

Due to our approximation of the spike train covariances (Eq. (23)), the dynamics of the mean synaptic weight within and across assemblies are coupled to each other only through correlations due to common inputs (the S_C terms).

Numerical solution of Eqs. (6) and (7) showed that training with $c_{AA} > 0$ and $c_{AB} = 0$ (20 minutes) produced robust assembly formation (p_{AA} increased, Figure 4B red lines) while maintaining low cross-assembly coupling (p_{AB} remained low, Figure 4C red lines). Furthermore, after training the assembly structure self-reinforced, with p_{AA} continuing to increase even though $c_{AA} = 0$ (Figure 4B,C outside the grey shaded area). These results are a main contribution of our study and represent a spike timing–based mechanism for assembly formation (Figure 1C) as an alternative to rate-based mechanisms (Figure 1B). We next used the low dimensionality of Eqs. (6) and (7) to analyze the dynamical mechanisms by which assembly formation occurred.

Similar to the network without assembly structure, the nullclines of p_{AA} and p_{AB} predicted their thresholds for potentiation. In order to numerically compute (for example) the p_{AA} nullcline, we found for each p_{AA} the p_{AB} that, in combination with the induced inhibitory weight p_{EI}^*, yielded $dp_{AA}/dt = 0$. Assuming that the eSTDP rule is temporally symmetric ($\tau_- \sim \tau_+ + \mathcal{O}(\epsilon)$) so that S_C and S_C^I both vanish permits an explicit calculation of the nullclines (Methods, 5.7):

$$p_\alpha^* = -\frac{\left(\bar{r}_E^2 S + c_\alpha \sigma^2 S_\eta\right) p_0}{\epsilon\left(S_F + p_0 S_B\right)} \tag{8}$$

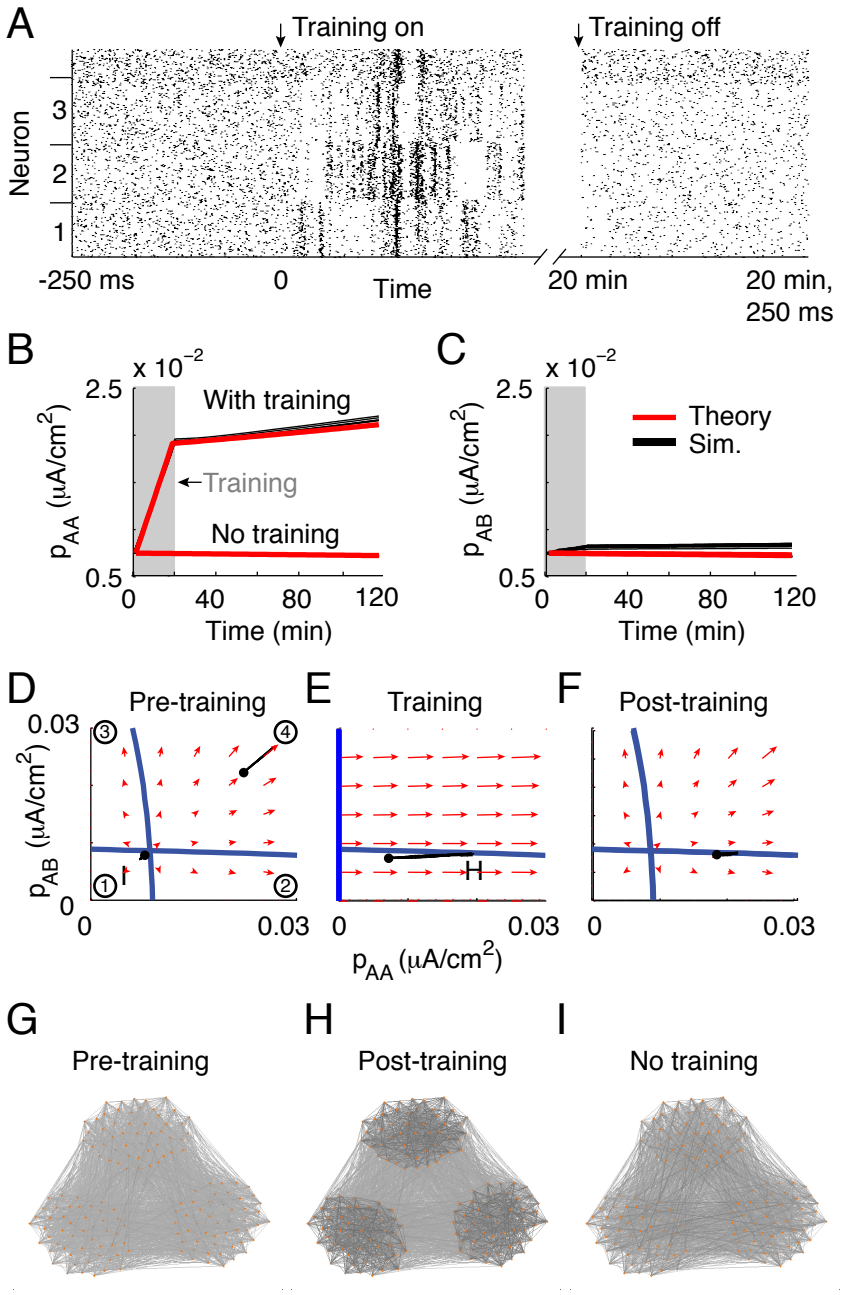

Figure 4. Training and spontaneous reinforcement of assembly structure. (A) Spike train raster of the network activity. Excitatory neurons are ordered by assembly membership (numbers on the ordinate axis). (B,C) Dynamics of the mean strength of within-assembly synapses (B) and of cross-assembly synapses (C). The red lines are from our theory in Eqs. (6) and (7). The black lines are computed from trial averaging the outputs spiking simulations. The grey boxes mark the stimulus duration. (D–F) Phase planes governing the mean synaptic weights before the training signal (D), during the training period (E) and after training (F). The numbers in D are referred to in the main text. The blue curves are nullclines computed from Eqs. (4), (6) and (7). (G–I) Visualization of a subset of the excitatory-excitatory connectivity. Nodes positioned by the Fruchterman-Reingold force algorithm.

for $\alpha \in \{AA, AB\}$. This then gives horizontal and vertical nullclines in (p_{AA}, p_{AB}) space, effectively decoupling the p_{AA} and p_{AB} dynamics. While this assumption is quantitatively inaccurate for our STDP rule (which has $\tau_- = 2\tau_+$), it reveals the main effect of external input correlations. Note that while the small parameter ϵ appears in the denominator, both terms of the numerator are also $\mathcal{O}(\epsilon)$ due to the balance between potentiation and depression in the eSTDP rule (i.e $S, S_\eta \sim \mathcal{O}(\epsilon)$).

Eq. (8) shows that p^*_α is positive in the absence of external input correlations ($c_\alpha = 0$). This is because we took $S < 0$ and the network is not fully connected ($p_0 < 1$), so that $S_F + p_0 S_B > 0$. In the absence of training, the fixed point (p^*_{AA}, p^*_{AB}) was unstable and the nullcline structure partitioned (p_{AA}, p_{AB}) space into four quadrants (Figure 4D): 1) a region where all structure dissolved because p_{AA} and p_{AB} both depressed, 2) a region where assembly structure formed since p_{AA} potentiated while p_{AB} depressed, 3) a region where a loop between assemblies formed because p_{AB} potentiated while p_{AA} depressed, and finally 4) a region where assemblies fused since p_{AA} and p_{AB} both potentiated. With $c_\alpha > 0$, the nullcline p^*_α was decreased by an amount proportional to c_α. In particular, $c_{AA} > 0$ reduced the threshold for potentiation of within-assembly connectivity, while leaving the threshold for cross-assembly connectivity unaffected. Thus, when an initial state was in region 1, training with $c_{AA} > 0$ and $c_{AB} = 0$ would result in the dynamics shifting to region 2 so that assembly structure formed (Figure 4D-F). Once training was completed, if p_{AA} increased sufficiently, then the state post-training remained in quadrant 2 and assembly structure continued to form, albeit at a slower rate. Thus an analysis of the mean field theory of Eqs. (6) and (7) gives a qualitative understanding of the dynamics of assembly formation.

We tested these mean field theory predictions in simulations of the full system of spiking neurons, divided into $M = 3$ assemblies. After 20 min of stimulation, we observed the formation of strongly connected assemblies of neurons (Figure 4G,H). The connectivity between assemblies was not potentiated; the assemblies did not fuse. We contrast this to the same network after 20 min of spontaneous activity: structure did not form spontaneously (Figure 4I). Furthermore, the mean field theory of Eqs. (6) and (7) gave an excellent match to the p_{AA} and p_{AB} estimated from the spiking network simulations (Figure 4B,C, black versus red curves). In total, our low-dimensional mean field theory not only gives a qualitative understanding of assembly formation through spike timing, but also gives a quantitatively accurate theory for the high-dimensional spiking network simulations upon which the theory is based.

Finally, while the synaptic strengths p_{AA} and p_{AB} evolved on a slow timescale of minutes, the STDP rule is sensitive to spike time correlations on a fast timescale of tens of milliseconds. Internally generated spike time correlations depend upon the recurrent network structure, and hence the covariance between neuron spike trains reflected the slow changes in p_{AA} and p_{AB}.

Indeed, spiking covariability after training was much larger within assemblies than between them (Figure 5A,B). Further, these differences were reinforced post-training, reflecting the concomitant dynamics of p_{AA} and p_{AB} during this time. Thus, the malleability of internal correlations provided a signature of assembly formation observable in the fast-timescale dynamics of coordinated spiking activity.

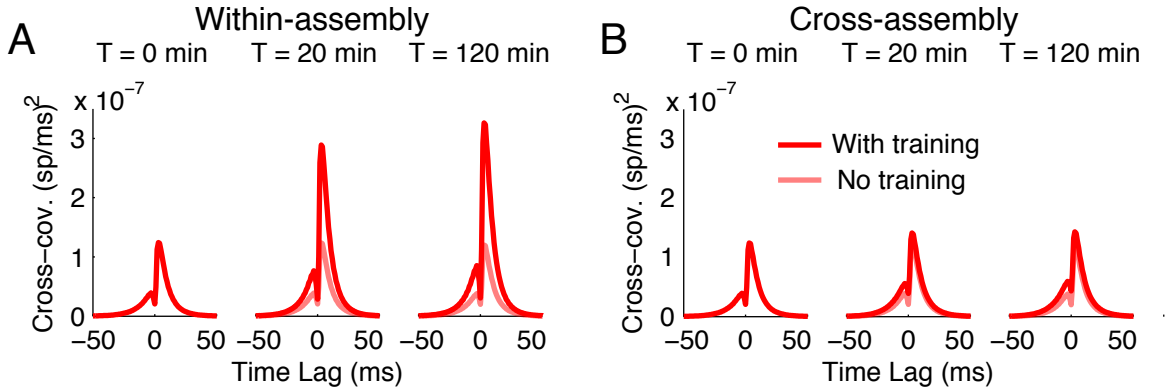

Figure 5. Spike train covariability reflects and reinforces learned network structure. (A) Average spike train cross-covariance between within-assembly pairs of neurons. (B) Average spike train cross-covariance between cross-assembly pairs of neurons. Cross-covariances estimated by the truncated linear response theory, Eq. (23). Solid: with training. Shaded: without training. Left: before training. Middle: end of stimulus presentation. Right: after spontaneous activity following stimulus presentation (as in Figure 4).

3.3 Reciprocal excitatory connectivity is preferentially promoted between similarly tuned neurons

In the previous section we examined how spatial correlations in external signals can promote the formation of neuronal assemblies. We discussed this only at the level of mean synaptic weights, the simplest measure of connectivity between neuron pairs. Recent data have revealed another striking feature of pair-based connectivity: pairs of neurons with similar stimulus preferences have strong reciprocal connectivity [7]. Theories of STDP focusing on pairs of neurons suggest that Hebbian STDP should suppress reciprocal connectivity [45, 54] (but see [16]). Our previous work has suggested that if reciprocal connectivity is sufficiently strong in a network on average, it can be reinforced by Hebbian STDP [38]. We next examined whether plasticity driven by precisely correlated spike times could contribute to the development of strong reciprocal connectivity in neuronal assemblies.

To that end we considered networks with partially symmetric baseline connectivity ($\Omega = 0.4$). This reciprocal structure is reflected in the weight matrix $\mathbf{W}$. To measure it in a way that allows us to take into account the development of stimulus-driven assemblies, we consider two metrics

of the network structure in addition to p_{AA} and p_{AB}:

$$\begin{aligned} \epsilon q_{AA} &= \frac{1}{\kappa^2} \sum_{i,j \in A} \mathbf{W}^0_{ij} \mathbf{W}_{ji} - p_0 \epsilon p_{AA} \\ \epsilon q_{AB} &= \frac{1}{\kappa \left(N_E - \kappa\right)} \sum_{i \in A} \sum_{j \notin A} \mathbf{W}^0_{ij} \mathbf{W}_{ji} - p_0 \epsilon p_{AB} \end{aligned} \tag{9}$$

These measure the average strength of reciprocal connections either within (q_{AA}) or between (q_{AB}) assemblies, above what would be expected by chance. As before, we assume symmetry between different assemblies. The inclusion of the mean reciprocal weights expands our description of the network structure to four dimensions ($p_{AA}, p_{AB}, q_{AA}, q_{AB}$). Furthermore, the dynamics of the mean synaptic weights p_{AA} and p_{AB}, in addition to depending on each other, now depend on q_{AA}, q_{AB} through the STDP-weighted covariances due to reciprocal connections (see Methods 5.6).

In order to obtain a simpler description, we considered the change of variables:

$$\begin{aligned} p_\Delta &= p_{AA} - p_{AB} \\ q_\Delta &= q_{AA} - q_{AB}. \end{aligned} \tag{10}$$

These measure the relative strength of assembly structure in the network, at the levels of mean connection strength (p_Δ) and above-chance reciprocal connection strength (q_Δ). In order for a network to respect the structure observed in mouse V1 by Cossell et al. [7], it should have $p_\Delta > 0, q_\Delta > 0$. The dynamics of (p_Δ, q_Δ) can be simply calculated from those of $(p_{AA}, p_{AB}, q_{AA}, q_{AB})$ and are (see Methods 5.8):

$$\frac{dp_\Delta}{dt} = c_\Delta \sigma^2 S_\eta p_0 + \epsilon \left[S_F p_\Delta + S_B (q_\Delta + p_0 p_\Delta) + S_C p_\Delta^2 \right] \tag{11}$$

$$\frac{dq_\Delta}{dt} = c_\Delta \sigma^2 S_\eta q_0 + \epsilon \left[S_F q_\Delta + S_B \left(1 - p_0\right) \left(q_\Delta + p_0 p_\Delta\right) + S_C \frac{q_0}{p_0} p_\Delta^2 \right] \tag{12}$$

where $c_\Delta = c_{AA} - c_{AB}$. Notably, the dynamics of p_Δ and q_Δ decoupled from the overall strengths of excitation and inhibition in the network, i.e., Eqs. (11) and (12) do not explictly depend on p_{AA}, p_{AB}, q_{AA} and q_{AB}. Further, the contribution of chance spike coincidences, $r_E^2 S$, canceled because neurons in each assembly have the same average firing rate. Satisfyingly, the mean field theory of Eqs. (11) and (12) gave an accurate match to network simulations during training ($c_\Delta > 0$) and spontaneous ($c_\Delta = 0$) regimes (Figure 6A,B).

Similar to the case of asymmetric networks, these dynamics admit nullclines that represent thresholds for potentiation/depression (Figure 6C, blue curves). The origin ($p_\Delta = 0, q_\Delta = 0$) is unstable and the nullclines divide the phase plane into four regions, containing each potential combination of potentiation and depression of (p_Δ, q_Δ). We take the synaptic weights to be

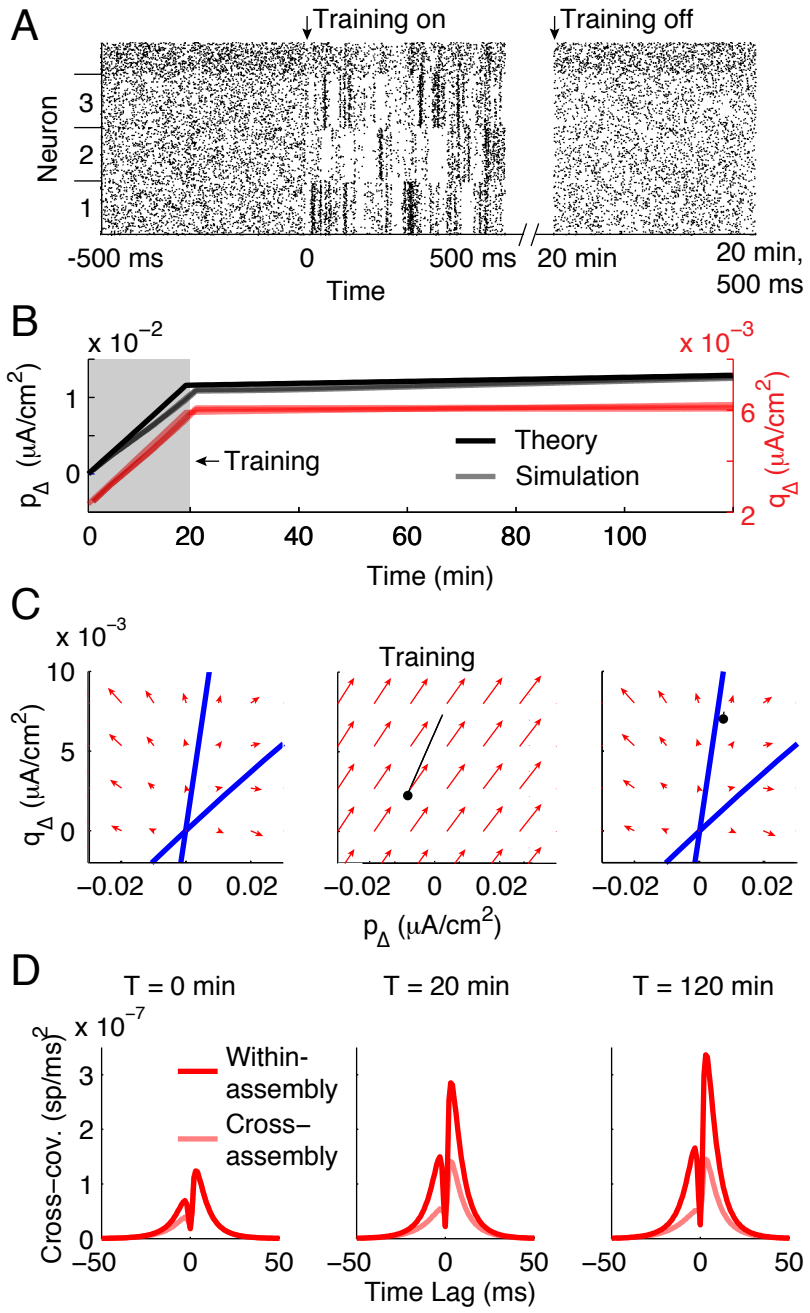

Figure 6. Reciprocal connectivity is preferentially promoted within assemblies. We consider networks with partially symmetric baseline connectivity. (A) Raster of network activity during pre-training, training and post-training phases. Excitatory neurons ordered by assembly membership (labeled on ordinate axis). (B) Time course of the relative strength of within-assembly synapses, p_Δ (black), and within-assembly reciprocal synapses, q_Δ (red). (C) Phase plane of p_Δ, q_Δ in the absence (pre- and post-training) or presence (training) of external input correlations. Blue: nullclines for $c_\Delta = 0$ were $p^*_\Delta = \frac{-(S_F+p_0S_B)\pm\sqrt{(S_F+p_0S_B)^2-4S_CS_Bq_\Delta}}{2S_C}$ and $q^*_\Delta = -\frac{S_B(1-p_0)p_0p_\Delta+S_C\frac{q_0}{p_0}p^2_\Delta}{S_F+S_B(1-p_0)}$. Black traces: simulation of the spiking network. (D) Average spike train cross-covariances (truncated approximation, Eq. (23)). Left, before training. Middle, immediately at end of training. Right, 100 min post-training.

initially unstructured, so that before training $p_\Delta \approx q_\Delta \approx 0$ (Figure 6C, left). If external input correlations are higher for within-assembly pairs than cross-assembly pairs ($c_\Delta > 0$), the unstable point at $(0, 0)$ is shifted to negative (p_Δ, q_Δ) (Figure 6C, middle). This pushed the unstable synaptic dynamics towards having assemblies of strongly reciprocally connected neurons (Figure 6C, right).

This shift in network structure was reflected by the magnitude of spike train covariances within and between assemblies. Indeed, the training of assembly structure into the network led to a doubling of spike train covariability for within-assembly neurons compared to cross-assembly neurons (Figure 6D). Due to the higher levels of reciprocal connectivity, the average spike train covariances at negative time lags were larger than for the network with $\Omega = 0$ (compare Figure 6D vs Figure 5). As was the case for asymmetric networks, these results suggest that spontaneously generated spike train correlations, in addition to providing a signature of learned network structure, can actively reinforce it.

3.4 Trained noise covariance maintains coding performance

We finally asked how the spontaneous reinforcement of learned network structures, and the associated internally generated spike train covariability, affected the ability of cell assemblies to encode their preferred inputs. We took the partially symmetric network (Figure 6) and allowed the external input to excitatory neurons in an assembly to depend on a stimulus θ: $\mu_{\mathrm{E}} = \mu_{\mathrm{ext}} + \mu_\theta$. For simplicity, we took each stimulus to target exactly one assembly and considered only the coding by a single assembly (labeled A; Figure 7A).

We measured the linear Fisher information [55] of an assembly's net activity $n_A = \sum_{i \in A} n_i$ about the stimulus θ:

$$FI_A = \left(\frac{dn_A}{d\theta} \right)^2 C_{AA}^{-1} \tag{13}$$

where n_A^i is the spike count (over $T = 100$ ms) from neuron i of assembly A, and C_{AA} is the variance of n_A. $dn_A/d\theta$ is the mean stimulus-response gain of the spike count of neurons in assembly A, so that $n_A = \kappa T r_A$ where r_A is the mean firing rate of a neuron in assembly A. Fisher information is a lower bound on the variance of any estimate of θ from r_A, and the restriction to linear Fisher information gives a natural decomposition of FI_A into the response gain in $dr_A/d\theta$ and response noise C_{AA}. Since r_A naively sums the assembly activity we have that for large N, the response variance $C_{AA} \propto \langle \mathrm{Cov}(n_A^i, n_A^j) \rangle_{ij}$, meaning that the mean pairwise covariance between neurons in an assembly is the dominate contribution to the noise in n_A's estimate of θ.

We compared stimulus coding in networks with and without trained network structure (Figure 7B top, black versus red curves). Since training increased spike train covariances for within-

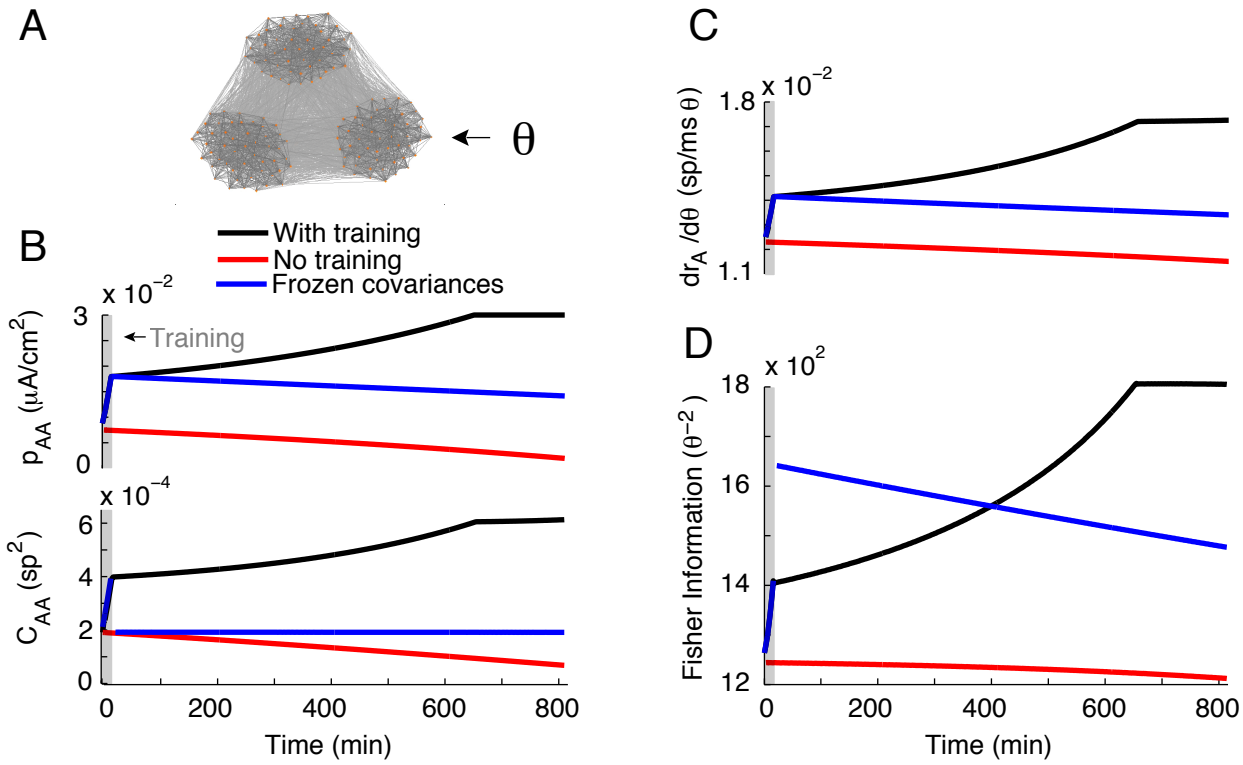

Figure 7. Spike train covariability reflects and reinforces learned network structure. (A) One assembly received a stimulus, θ, which it encodes by that assembly's total spike count in $T = 100$ ms. (B) Top: Variance of the summed spike count of assembly A increased during and after training (black). Assembly A's spike count variance decreased without training (red). As a control, we reset spiking covariability after training to pre-training values and froze them (blue). Bottom: The gain of stimulated neurons with respect to. θ increased during and after training (black). Without training, the stimulus-response gain decreased (red). With frozen covariances, the gain decreased after training (blue). Grey box: training period. (C) The mean strength of within-assembly connectivity. (D) Fisher information of the assembly's spike count about θ.

assembly pairs (Figure 6D) networks, then the variance of the assembly's summed spike count also increased with training and was reinforced after training (Figure 7B bottom, black curve). In agreement, when the training signals were absent and assemblies did not form, then C_{AA} slowly decreased over time (Figure 7B bottom, red curve). One expectation from increased variability is that training assembly structure would, in our simplified coding scenario, be deleterious to stimulus coding (since $FI_A \propto C_{AA}^{-1}$).

To determine the net impact of training on FI_A, we first evaluated how training and covariability affected response gain, to then be combined with C_{AA} to ultimately yield FI_A. We calculated the stimulus-response gain $dn_A/d\theta$, taking into account direct stimulus-driven inputs and indirect filtering of the stimulus through recurrence onto assembly A (see Methods 5.9). A consequence of trained assembly structure was increased gain through the positive feedback inherent within an assembly (Figure 7C, black versus red curves). The increased gain outweighed the increased variability so that overall FI_A grew with training and further increased through the assembly

reinforcement post-training (Figure 7D, black versus red curves). In total, assembly formation was overall beneficial to network coding despite requiring larger overall network variability.

As a final illustration of the beneficial role of noise correlations for stimulus coding, we considered an artificial network where training occurred, yet immediately after training we reset the spike train covariability to its pre-training value and forced it to remain at this value (Figure 7B, blue curves). For a period of time post-training the FI_A from this network was larger than that of the trained network without a reseted covariability, owing to the combination of a large gain from training and low variability through the artificial reset (Figure 7D, blue versus black curves). However, a consequence of low spike train covariability was a slow but clear degradation of assembly structure so that response gain reduced over time. This eventually reduced FI_A so that the network with internally generated covariability showed higher FI_A for times > 400 minutes after training. Thus, while noise correlations can have a detrimental impact on stimulus coding, the benefits of stimulus-specific recurrent structure and the role of spike train correlations play in maintaining that structure are such that noise correlations were beneficial in our simplified coding scenario.

4 Discussion

Theoretical work with eSTDP in cortical networks first established the role of timing in the formation of feedforward structures [14–17,20]. More recently, eSTDP has been shown to promote the spontaneous formation of structured circuit motifs [38] as well as support the stability of attractor network structure [56]. However, the role of spike timing in the formation of trained macroscopic assembly structure has been elusive. We derived a low-dimensional mean field theory for the plasticity of neuronal assemblies in partially symmetric networks of integrate-and-fire neurons with excitatory STDP and homeostatic inhibitory STDP. This revealed that internally generated spike train correlations can provide a threshold for potentiation or depression of mean synaptic weight and for mean reciprocal connectivity. Spatial correlations in external inputs shifted these thresholds, promoting an assembly structure in the network. Furthermore, the post-training structure of spike train correlations reflected the learned network structure and actively reinforced the architecture. This promoted strong synaptic weights within assemblies and strong reciprocal connectivity within assemblies.

4.1 Rate-based versus timing–based assembly formation

Since early seminal work [43, 44, 57] there has been intense research in the role of spike timing in shaping synaptic strength [2, 13]. While much theoretical work focused on phenomenological eSTDP plasticity rules (like the one used in our study) [40, 45, 54], there have also been advances in biophysically based models of eSTDP [25, 26, 58, 59]. These realistic models capture the known firing rate–dependence of eSTDP [27], complicating the discussion surrounding the role of spike

timing in synaptic learning. Indeed, past work in recurrently coupled networks of spiking neuron models has shown that the rate-dependence of these models can be sufficient for forming and maintaining neuronal assembly structure [22, 23]. Our study gives an alternative framework, where it is the fine-timescale correlations in spiking activity that drives assembly formation.

The mechanisms behind rate-based and timing–based assembly formation are distinct. In the rate-based scenario, the training of assemblies is *sequential*—each stimulus is presented in the absence of other stimuli so that neuron pairs within the same assembly can have coordinated high firing rates to drive potentiation, while neuron pairs in different assemblies can have a high-low firing rates that drive depression. By contrast, in the timing-based framework, assemblies can be trained in *parallel* since within- and cross-assembly neuron pairs can simultaneously receive correlated and uncorrelated external inputs. Further, while both frameworks show a spontaneous reinforcement of assembly structure, the mechanics of reinforcement are quite different. In rate-based assembly formation the learned network structure is a *stable* attractor in the space of synaptic weights. Spontaneous reinforcement occurs if the network has not converged to the attractor during training [22, 23]. If the network structure is perturbed from the attractor, then spontaneous activity will retrain the network [22]. In our timing-based formation, it is the position of an unstable *repeller* that determines the growth or decay of structure. Spontaneous reinforcement occurs when the synaptic state is such that the repeller pushes dynamics towards more structured assembly wiring.

Unstable solutions in synaptic learning are a reflection of a competitive synaptic interaction often associated with additive Hebbian STDP rules [45,46]. Past studies have used this instability to drive feedforward structure [14, 15, 17, 20], in effect harnessing the causality-rewarding nature of the Hebbian rule. Our work shows that this competitive synaptic dynamic can also be used to drive assembly structure. At the surface this seems counterintuitive since assembly dynamics are thought of as cooperative (within the assembly). In our model this is misleading and the competition between different subcomponents of spike train covariability (forwards, backwards and common synaptic wiring) supports robust assembly wiring. A recent study that uses a similar theoretical framework to ours has also shown how balanced STDP ($S \sim \mathcal{O}(\epsilon)$) can support the spontaneous emergence of assembly structure in small networks [16]. However, the STDP rule of that study was acausal, with near coincident spikes strengthening both forward and backward connections. In that study the dynamics of assembly formation did not rely on competitive synaptic interactions and there is no relationship to external stimuli; it is thus quite distinct from that exposed in our study.

While the mechanisms underlying rate- and timing-based assembly formation via eSTDP are distinct, that is not to say that they are mutually exclusive. Synaptic plasticity clearly has rate- and timing-based components and both can actively reinforce assembly structure, suggesting that

the mechanisms may be cooperative. Future work should investigate their interactions during learning.

4.2 Inhibitory plasticity and inhibitory stabilization

Inhibitory feedback plays two main roles in this study. The first is to modulate excitatory plasticity by contributing to spike train covariability amongst excitatory neurons. The strength of this contribution is governed by the strength of the inhibitory feedback, which is in turn governed by inhibition's second role: homeostatic control of firing rates. Inhibition's role in stabilizing network activity in the face of strong recurrent excitation has been the focus of much recent work in theoretical neuroscience. Notably, strong inhibitory feedback provides dynamical explanations for the generation of variable and asynchronous activity [60–63] and can also account for paradoxical responses to external inhibitory inputs [64] and diverse features of tuning in visual cortex [65].

In the absence of inhibition, potentiation of excitatory synapses in our networks led to runaway excitation, meaning that in the presence of inhibition the network existed in an inhibitory-stabilized regime (Figure 3). In contrast to other recent studies [22, 23], inhibitory STDP alone was sufficient to stabilize the network activity in our work without imposing synaptic scaling or other compensatory mechanisms. This was due to the relationship between the timescales of excitatory and inhibitory plasticity. We take the excitatory plasticity to be balanced between potentiation and depression. This sets the dynamics of the mean excitatory weight p to occur on an $\mathcal{O}(\epsilon^{-1})$ timescale set by the eSTDP rule and the magnitude of spike train correlations. When the firing rates are maintained at their stable fixed points, the inhibitory STDP is similarly governed by a timescale set by the iSTDP rule and the magnitude of spike train correlations. If the firing rates are outside an $\mathcal{O}(\epsilon)$ neigborhood around their fixed point, this causes the iSTDP rule to become unbalanced, so that it is governed by an $\mathcal{O}(1)$ timescale (Methods, 5.5). This feature—that the inhibitory STDP can become unbalanced in order to maintain stable activity—guarantees that it can dynamically stabilize the network activity in the face of the balanced excitatory plasticity.

The question of how neurons can undergo associative, Hebbian learning while maintaining stable activity has long been studied [66, 67]. While homeostasis is often thought of as a slower process than learning, recent work has highlighted the necessity of homeostatic mechanisms operating on a comparable timescale to excitatory plasticity [68]. Homeostatic regulation acting alone, however, can paradoxically destabilize network activity, inducing oscillations in neurons' firing rates [69]. Homeostatic regulation mediated by diffusive transmitters like nitrous oxide can have different effects than that mediated by synaptic mechanisms [70]. The study of how homeostatic regulation and mechanisms for associative learning interact to allow stable memories and stable activity remains an exciting area of open inquiry.

4.3 Correlated spontaneous activity can maintain coding performance

Many theoretical studies have asked how the joint trial-to-trial fluctuations in population response (noise correlations) impact population coding [71–73]. The answer to this question depends on many factors. In particular, the impact of noise correlations on coding depends on how they relate to neurons' stimulus preferences [74–81]. In our study we explore a novel and complementary viewpoint on the impact of noise correlations on population coding.

Noise correlations are often related to neurons' stimulus tuning [31, 82–85]—neuron pairs with similar tuning show larger noise correlations than pairs with dissimilar tuning. The mechanisms behind noise correlations are varied [86], and both feedforward [87] and recurrent [63] circuits can contribute to linking stimulus and noise correlations. Further, in the absence of sensory stimulation, patterns of activity across cortical populations are often similar to those observed during sensory stimulation [88–94]. Thus, the circuits that support correlated variability in spontaneous states likely overlap with the circuits responsible for noise correlations in evoked states. In other words, noise correlations may simply be a reflection of circuit dynamics that occur during periods when stimulus coding is not being performed.

Spontaneous activity is usually viewed as a problem for plasticity: learned weight changes must be stable in the face of spontaneous activity. Some previous studies have addressed this issue by endowing individual synapses with dynamical bistability between weak and strong weights [23, 26, 95, 96]. By contrast, in our study the trained network architecture produced sizable within-assembly spontaneous correlations that combined with the STDP rule to reinforce assembly structure. Thus, spontaneous activity did not dissolve learned architecture but rather preserved trained wiring. If the assembly wiring was originally due to a shared stimulus input, then the spontaneous correlations needed to retain structure will be a source of noise correlations when the stimulus is to be coded.

Our simplified stimulus coding scenario was such that within-assembly noise correlations degraded the neural code. However, the strong positive feedback from within-assembly recurrence enhanced the response gain, which improved coding. Many studies of population coding separate response gain and response variability and for the purposes of analysis they are conceived as independent from one another [74, 75, 77]. While these studies have given insight into population coding, relating noise correlations and response gain to one another complicates analysis significantly [76, 80, 81, 97]. Our study expands on this general idea so that noise correlations are a reflection of the active maintenance of assembly structure and the high response gain it confers to a neuronal population. This finding does not critically depend on the fast-timescale coordinated spiking activity required for STDP, and stability of assembly structure through long-timescale firing rate correlations should have a similar effect [22, 56].

5 Methods

5.1 Network model

We consider a network of N neurons, N_E of which are excitatory and divided into M clusters of size κ. There are $N_I = \gamma\kappa$ inhibitory neurons. Model parameters are in Table 1. We take the excitatory-excitatory block of the adjacency matrix $\mathbf{W}^0$ to be partially symmetric: $\mathbf{W}^0_{EE} = (p_0 - p_{0\text{sym}})\,\mathbf{W}^0_{\text{ER}} + p_{0\text{sym}}\mathbf{W}^0_{\text{sym}}$ where $\mathbf{W}^0_{\text{ER}}$ has (directed, i.e. non-symmetric) Erdős-Rényi statistics and $\mathbf{W}^0_{\text{sym}}$ is symmetric with Erdős-Rényi statistics (i.e., as in an undirected graph). Additionally we exclude autapses ($\mathbf{W}^0_{ii} = 0\ \forall i \in 1, \ldots, N$).

This means that the excitatory-excitatory connectivity is characterized by its empirical connection density p_0 and the frequency of loops q_0

$$
\begin{aligned}
p_0 &= \frac{1}{N_E^2}\sum_{i,j=1}^{N_E} \mathbf{W}^0_{ij} \\
q_0 &= \frac{1}{N_E^2}\sum_{i,j=1}^{N_E} \mathbf{W}^0_{ij}\mathbf{W}^0_{ji} - p_0^2
\end{aligned}
\tag{14}
$$

Table 1. Model parameters

Parameter	Description	Value
C	Membrane capacitance	1 μF/cm^2
g_L	Leak conductance	0.1mS/cm^2
V_L	Leak reversal potential	-72 mV
Δ	Action potential steepness	1.4 mV
V_T	Action potential initiation threshold	-48 mV
V_{th}	Action potential threshold	30 mV
V_{re}	Action potential reset	-72 mV
τ_{ref}	Action potential width	2 ms
μ	External input mean	1 μA/cm^2
σ	External input standard deviation	9 mV
$W^{\text{max,E}}$	Maximum synaptic weight	15ϵ μA/cm^2
$W^{\text{max,I}}$	Maximum synaptic weight	-7.5ϵ μA/cm^2
$J(t)$	Synaptic filter (EPSC shape)	$\exp -(t/\tau_s)$
τ_{sE}	Excitatory synaptic time constant	2 ms
τ_{sI}	Inhibitory time constant	10 ms

We assume that the statistics of the adjacency matrix for within- and between-assembly connectivity are the same (and equal to p_0 and q_0). The synaptic weight matrix, $\mathbf{W}$, is initially

generated from $\mathbf{W}^0$ by giving each synapse the same initial weight. We consider the mean strength of E-E synapses within one cluster A and from other clusters into cluster A, p_{AA} and p_{AB} respectively:

$$\begin{aligned} \epsilon p_{AA} &= \frac{1}{\kappa^2} \sum_{i,j \in A} \mathbf{W}_{ij} \\ \epsilon p_{AB} &= \frac{1}{\kappa(N_E - \kappa)} \sum_{i \in A} \sum_{j \notin A} \mathbf{W}_{ij} \end{aligned} \tag{15}$$

The small parameter $\epsilon = (\kappa p_0)^{-1}$ scales the synaptic weights. We take the mean strength of connections within each cluster to be symmetric and the strength of connections into any one cluster from outside to be the same as into the others (so for all clusters A and B, $p_{AA} = p_{BB}$ and $p_{AB} = p_{BA}$). Similarly, we measure the strength of *reciprocal* connections within a cluster, q_{AA}, or between clusters, q_{AB}:

$$\begin{aligned} \epsilon q_{AA} &= \frac{1}{\kappa^2} \sum_{i,j \in A} \mathbf{W}_{ij} \mathbf{W}^0_{ji} - \epsilon p_0 p_{AA} \\ \epsilon q_{AB} &= \frac{1}{\kappa(N_E - \kappa)} \sum_{i \in A} \sum_{j \notin A} \mathbf{W}_{ij} \mathbf{W}^0_{ji} - \epsilon p_0 p_{AB} \end{aligned} \tag{16}$$

By subtracting off $p_0 p_{AA}$ in the definition of q_{AA} (and likewise for q_{AB}), we measure the mean strength of reciprocal connections above what would be expected in a network with no correlations between synapses. Note: if the network is asymmetric ($\mathbf{W}^0_{\text{sym}} = \mathbf{0}$) then q_0 is negligible ($\mathcal{O}(\epsilon^{-3/2})$) and so are the initial values of q_{AA} and q_{AB}.

We take the connectivity in between inhibitory and excitatory neurons, and within inhibitory neurons, to have (asymmetric) Erdős-Rényi statistics, so that these are characterized by their mean synaptic weights: p_{EI} for inhibitory $\rightarrow$ excitatory connections,

$$\epsilon p_{EI} = \frac{1}{N_E N_I} \sum_{i=1}^{N_E} \sum_{j=N_E+1}^{N} \mathbf{W}_{ij}, \tag{17}$$

and likewise p_{IE} and p_{II}.

Finally, individual neurons had exponential integrate-and-fire (EIF) dynamics [42], part of a class of models well-known to capture the spike initiation dynamics of cortical neurons [98, 99]. Neurons' membrane voltages obeyed:

$$C \frac{dV_i}{dt} = g_L \left(V_L - V_i\right) + g_L \Delta \exp\left(\frac{V_i - V_T}{\Delta}\right) + I_{\text{i}}(t) + \sum_{j=1}^{N} \mathbf{W}_{ij} \left(J(t) * y_j(t).\right). \tag{18}$$

with model parameters in Table 1.

5.2 Plasticity models

Synapses between excitatory neurons undergo additive Hebbian STDP:

$$\epsilon L(s) = \begin{cases} \mathcal{H}(W^{\max} - \mathbf{W}_{ij}) f_+ e^{-\frac{|s|}{\tau_+}}, & \text{if } s \geq 0 \\ \mathcal{H}(\mathbf{W}_{ij})\,(-f_-)\, e^{-\frac{|s|}{\tau_-}}, & \text{if } s < 0, \end{cases} . \tag{19}$$

where $s = t_{post} - t_{pre}$ is the time lag between spikes. $f_\pm$ give the amplitude of individual changes in synaptic weights due to potentiation (f_+) or depression (f_-), and the time constants $\tau_\pm$ determine how synchronous spike pairs must be to cause plasticity. When $f_\pm \ll W^{\max}$, so that the timescale of plasticity is much longer than that of the STDP rule, individual weights undergo diffusion [40] and their drift can be calculated as:

$$\frac{d\mathbf{W}_{ij}}{dt} = \mathbf{W}^0_{ij} \int_{-\infty}^{\infty} \epsilon L(s)\big(r_i r_j + \mathbf{C}_{ij}(s)\big) ds. \tag{20}$$

Here, r_i is the time-averaged firing rate of neuron i and $\mathbf{C}_{ij}(s)$ is the spike train cross-covariance function of neurons i and j. We will assume that the integral of $L(s)$ is small enough ($\mathcal{O}(\epsilon)$) so that firing rates do not dominate the plasticity.

The inhibitory STDP rule is

$$\epsilon L_I(s) = \mathcal{H}(\mathbf{W}_{ij} - W^{\max,\mathrm{I}}) f_I e^{-\frac{|s|}{\tau_I}}. \tag{21}$$

In addition to this pair-based rule, each presynaptic (inhibitory) spike drives depression of the inhibitory synapses by $\mathcal{H}(-\mathbf{W}_{ij}) d_I = -2 f_I \bar{r}_E \tau_I$. This gives inhibitory $\rightarrow$ excitatory synapses a drift of

$$\frac{d\mathbf{W}_{ij}}{dt} = \mathbf{W}^0_{ij}\Big(\int_{-\infty}^{\infty} \epsilon L_I(s)\big(r_i r_j + \mathbf{C}_{ij}(s)\big) ds - 2 f_I \tau_I \bar{r}_E r_j \Big). \tag{22}$$

5.3 Derivation of assembly dynamics

Here we will derive the dynamics of the assembly structure in networks of integrate-and-fire neurons undergoing STDP. We will begin by considering the dynamics of mean synaptic weights and mean reciprocal synaptic weights both within and between assemblies. The dynamics of (p, q) and (p_Δ, q_Δ) considered in the main text will then be recovered at the end. The derivation follows the same steps as the derivation of the motif dynamics in [38]. We begin by expanding the covariance matrix $\mathbf{C}$ in path lengths through the network [39, 100] and truncating at first

order in the interactions to obtain:

$$\mathbf{C}_{ij}(s) \approx \overbrace{\delta_{ij}\mathbf{C}^0_{ij}(s)}^{\text{autocovariance}} + \overbrace{\left(\mathbf{A}_i * \mathbf{C}^\eta * \mathbf{A}_j\right)(s)}^{\text{external inputs}} + \overbrace{\left(\mathbf{W}_{ij}\mathbf{K}_{ij} * \mathbf{C}^0_{jj}\right)(s)}^{\text{forwards connections}} + \overbrace{\left(\mathbf{C}^0_{ii} * \mathbf{W}_{ji}\mathbf{K}^-_{ji}\right)(s)}^{\text{backwards connections}}$$
$$+ \underbrace{\sum_{k=1}^{N_E}\left(\mathbf{W}_{ik}\mathbf{K}_{ik} * \mathbf{C}^0_{kk} * \mathbf{W}_{jk}\mathbf{K}^-_{jk}\right)(s)}_{\text{common E inputs}} + \underbrace{\sum_{k=N_E+1}^{N}\left(\mathbf{W}_{ik}\mathbf{K}_{ik} * \mathbf{C}^0_{kk} * \mathbf{W}_{jk}\mathbf{K}^-_{jk}\right)(s)}_{\text{common I inputs}}. \tag{23}$$

As can be seen from Eqs. (20) and (22), these cross-covariances will control plasticity through their integral against the STDP rule. We define variables measuring these STDP-weighted covariances (factoring out their amplitude, given by the $\mathbf{W}_{ij}$ factors in Eq. (23)):

$$\begin{aligned}
S &= \int_{-\infty}^{\infty} L(s)ds \\
S_\eta &= \int_{-\infty}^{\infty} L(s)\left(A_E(t) * A_E(-t)\right)ds \\
S_F &= \int_{-\infty}^{\infty} L(s)\left(K_{EE}(t) * C^0_E(s)\right)ds \\
S_B &= \int_{-\infty}^{\infty} L(s)\left(C^0_E(s) * K_{EE}(-t)\right)ds \\
S_C &= \int_{-\infty}^{\infty} L(s)\left(K_{EE}(t) * C^0_E(s) * K_{EE}(-t)\right)ds \\
S^I_C &= \int_{-\infty}^{\infty} L(s)\left(K_{EI}(t) * C^0_I(s) * K_{EI}(-t)\right)ds
\end{aligned} \tag{24}$$

and

$$\begin{aligned}
S^I &= \int_{-\infty}^{\infty} L_I(s)ds = 2f_I\tau_I \\
S^{EI}_\eta &= \int_{-\infty}^{\infty} L_I(s)\big(A_E(t) * A_I(-t)\big)ds \\
S^{EI}_F &= \int_{-\infty}^{\infty} L_I(s)\big(K_{EI}(t) * C^0_I(s)\big)ds \\
S^{EI}_B &= \int_{-\infty}^{\infty} L_I(s)\big(C^0_E(s) * K_{IE}(-t)\big)ds \\
S^{EIE}_C &= \int_{-\infty}^{\infty} L_I(s)\big(K_{EE}(t) * C^0_E(s) * K_{IE}(-t)\big)ds \\
S^{EII}_C &= \int_{-\infty}^{\infty} L_I(s)\big(K_{EI}(t) * C^0_I(s) * K_{II}(-t)\big)ds.
\end{aligned} \tag{25}$$

In each of these definitions, $A_\alpha(t)$ corresponds to the mean linear response function of neurons of type α, $\alpha \in \{E, I\}$. $K_{\alpha\beta}(t)$ is the convolution of $A_\alpha(t)$ and the synaptic filter for synapses from β neurons to α neurons ($\alpha, \beta \in \{E, I\}$). We also define r_E and r_I, the average excitatory and inhibitory firing rates. Note that each of these are implicitly functions of the mean synaptic

drive onto excitatory and inhibitory neurons. Note that for the iSTDP rule, each presynaptic spike causes depression by $-S^I \bar{r}_E$.

We want the dynamics of the connectivity variables $p_{AA}, p_{AB}, q_{AA}, q_{AB}$, so we differentiate these with respect to time. Then, inserting Eq. (23) into Eq. (20) and this into dp_{AA}/dt yields:

$$\begin{aligned}\frac{dp_{AA}}{dt} &= \left(r_E^2 S + c_{AA}\sigma^2 S_\eta\right)\frac{1}{\kappa^2}\sum_{i,j\in A}\mathbf{W}_{ij}^0 + S_F\frac{1}{\kappa^2}\sum_{i,j\in A}\mathbf{W}_{ij}^0\mathbf{W}_{ij} + S_B\frac{1}{\kappa^2}\sum_{i,j\in A}\mathbf{W}_{ij}^0\mathbf{W}_{ji} \\ &+ S_C\frac{1}{\kappa^2}\sum_{i,j\in A}\sum_{k=1}^{N_E}\mathbf{W}_{ij}^0\mathbf{W}_{ik}\mathbf{W}_{jk} + S_C^I\frac{1}{\kappa^2}\sum_{i,j\in A}\sum_{k=N_E+1}^{N}\mathbf{W}_{ij}^0\mathbf{W}_{ik}\mathbf{W}_{jk}\end{aligned} \tag{26}$$

and similar for p_{AB}:

$$\begin{aligned}\frac{dp_{AB}}{dt} &= \left(r_E^2 S + c_{AB}\sigma^2 S_\eta\right)\frac{1}{\kappa(N_E-\kappa)}\sum_{i\in A}\sum_{j\notin A}\mathbf{W}_{ij}^0 + S_F\frac{1}{\kappa(N_E-\kappa)}\sum_{i\in A}\sum_{j\notin A}\mathbf{W}_{ij}^0\mathbf{W}_{ij} \\ &+ S_B\frac{1}{\kappa(N_E-\kappa)}\sum_{i\in A}\sum_{j\notin A}\mathbf{W}_{ij}^0\mathbf{W}_{ji} + S_C\frac{1}{\kappa(N_E-\kappa)}\sum_{i\in A}\sum_{j\notin A}\sum_{k=1}^{N_E}\mathbf{W}_{ij}^0\mathbf{W}_{ik}\mathbf{W}_{jk} \\ &+ S_C^I\frac{1}{\kappa(N_E-\kappa)}\sum_{i\in A}\sum_{j\notin A}\sum_{k=N_E+1}^{N}\mathbf{W}_{ij}^0\mathbf{W}_{ik}\mathbf{W}_{jk}.\end{aligned} \tag{27}$$

The mean bidirectional connection strengths similarly evolve according to:

$$\begin{aligned}\frac{dq_{AA}}{dt} &= \left(r_E^2 S + c_{AA}\sigma^2 S_\eta\right)\frac{1}{\kappa^2}\sum_{i,j\in A}\mathbf{W}_{ij}^0\mathbf{W}_{ji}^0 + S_F\frac{1}{\kappa^2}\sum_{i,j\in A}\mathbf{W}_{ij}^0\mathbf{W}_{ij}\mathbf{W}_{ji}^0 + S_B\frac{1}{\kappa^2}\sum_{i,j\in A}\mathbf{W}_{ij}^0\mathbf{W}_{ji}\mathbf{W}_{ji}^0 \\ &+ S_C\frac{1}{\kappa^2}\sum_{i,j\in A}\sum_{k=1}^{N_E}\mathbf{W}_{ij}^0\mathbf{W}_{ik}\mathbf{W}_{jk}\mathbf{W}_{ji}^0 + S_C^I\frac{1}{\kappa^2}\sum_{i,j\in A}\sum_{k=N_E+1}^{N}\mathbf{W}_{ij}^0\mathbf{W}_{ik}\mathbf{W}_{jk}\mathbf{W}_{ji}^0 - p_0\frac{dp_{AA}}{dt}\end{aligned} \tag{28}$$

$$\begin{aligned}\frac{dq_{AB}}{dt} &= \left(r_E^2 S + c_{AB}\sigma^2 S_\eta\right)\frac{1}{\kappa(N_E-\kappa)}\sum_{i\in A}\sum_{j\notin A}\mathbf{W}_{ij}^0\mathbf{W}_{ji}^0 + S_F\frac{1}{\kappa(N_E-\kappa)}\sum_{i\in A}\sum_{j\notin A}\mathbf{W}_{ij}^0\mathbf{W}_{ij}\mathbf{W}_{ji}^0 \\ &+ S_B\frac{1}{\kappa(N_E-\kappa)}\sum_{i\in A}\sum_{j\notin A}\mathbf{W}_{ij}^0\mathbf{W}_{ji}\mathbf{W}_{ji}^0 + S_C\frac{1}{\kappa(N_E-\kappa)}\sum_{i\in A}\sum_{j\notin A}\sum_{k=1}^{N_E}\mathbf{W}_{ij}^0\mathbf{W}_{ik}\mathbf{W}_{jk}\mathbf{W}_{ji}^0 \\ &+ S_C^I\frac{1}{\kappa(N_E-\kappa)}\sum_{i\in A}\sum_{j\notin A}\sum_{k=N_E+1}^{N}\mathbf{W}_{ij}^0\mathbf{W}_{ik}\mathbf{W}_{jk}\mathbf{W}_{ji}^0 - p_0\frac{dp_{AB}}{dt}.\end{aligned} \tag{29}$$

The mean inhibitory-to-excitatory synaptic weight obeys:

$$\begin{aligned}\frac{dp_{EI}}{dt} &= \frac{1}{N_E N_I}\sum_{i=1}^{N_E}\sum_{j=N_E+1}^{N}\frac{d\mathbf{W}_{ij}}{dt}\\ &= \frac{1}{N_E N_I}\sum_{i=1}^{N_E}\sum_{j=N_E+1}^{N}\mathbf{W}^0_{ij}\Big(\int_{-\infty}^{\infty}\epsilon L_I(s)\big(r_i r_j + \mathbf{C}_{ij}(s)\big)ds - 2f_I\tau_I\bar{r}_E r_I\Big).\end{aligned}$$

Inserting the first-order truncation of spike train covariances yields:

$$\begin{aligned}\frac{dp_{EI}}{dt} &= \big(r_I\big(r_E-\bar{r}_E\big)S^I + c_{EI}\sigma^2 S^{EI}_\eta\big)\frac{1}{N_E N_I}\sum_{i=1}^{N_E}\sum_{j=N_E+1}^{N}\mathbf{W}^0_{ij}\\ &+ S^{EI}_F\frac{1}{N_E N_I}\sum_{i=1}^{N_E}\sum_{j=N_E+1}^{N}\mathbf{W}^0_{ij}\mathbf{W}_{ij} + S^{EI}_B\frac{1}{N_E N_I}\sum_{i=1}^{N_E}\sum_{j=N_E+1}^{N}\mathbf{W}^0_{ij}\mathbf{W}_{ji}\\ &+ S^{EIE}_C\frac{1}{N_E N_I}\sum_{i=1}^{N_E}\sum_{j=N_E+1}^{N}\sum_{k=1}^{N_E}\mathbf{W}^0_{ij}\mathbf{W}_{ik}\mathbf{W}_{jk} + S^{EII}_C\frac{1}{N_E N_I}\sum_{i=1}^{N_E}\sum_{j=N_E+1}^{N}\sum_{k=N_E+1}^{N}\mathbf{W}^0_{ij}\mathbf{W}_{ik}\mathbf{W}_{jk}.\end{aligned} \tag{30}$$

The next step in writing down dynamics for each of the p and q variables of interest is to evaluate the sums over $\mathbf{W}$ and $\mathbf{W}^0$ in Eqs. (26)–(30). Recalling that the adjacency matrix is Erdős-Rényi except for the partial symmetry of the excitatory-excitatory block, this yields (neglecting higher-order motif contributions):

$$\frac{dp_{AA}}{dt} = \big(r_E^2 S + c_{AA}\sigma^2 S_\eta\big)p_0 + S_F\epsilon p_{AA} + S_B\epsilon\left(q_{AA}+p_0 p_{AA}\right) + S_C\epsilon^2 p_0\left(\kappa p_{AA}^2 + (N_E-\kappa)\,p_{AB}^2\right) + S_C^I\epsilon^2 N_I p_0 p_{EI}^2 \tag{31}$$

$$\begin{aligned}\frac{dp_{AB}}{dt} &= \big(r_E^2 S + c_{AB}\sigma^2 S_\eta\big)p_0 + S_F\epsilon p_{AB} + S_B\epsilon(q_{AB}+p_0 p_{AB})\\ &\quad + S_C\epsilon^2 p_0\left(2\kappa p_{AA}p_{AB} + (N_E-2\kappa)\,p_{AB}^2\right) + S_C^I\epsilon^2 N_I p_0 p_{EI}^2\end{aligned} \tag{32}$$

$$\begin{aligned}\frac{dq_{AA}}{dt} &= \big(r_E^2 S + c_{AA}\sigma^2 S_\eta\big)q_0 + S_F\epsilon q_{AA} + S_B\epsilon(1-p_0)(q_{AA}+p_0 p_{AA})\\ &\quad + S_C\epsilon^2 q_0\left(\kappa p_{AA}^2 + (N_E-\kappa)p_{AB}^2\right) + S_C^I\epsilon^2 N_I q_0 p_{EI}^2\end{aligned} \tag{33}$$

$$\begin{aligned}\frac{dq_{AB}}{dt} &= \big(r_E^2 S + c_{AB}\sigma^2 S_\eta\big)q_0 + S_F\epsilon q_{AB} + S_B\epsilon\left(1-p_0\right)\left(q_{AB}+p_0 p_{AB}\right)\\ &\quad + S_C\epsilon^2 q_0\left(2\kappa p_{AA}p_{AB} + (N_E-2\kappa)\,p_{AB}^2\right) + S_C^I\epsilon^2 q_0 N_I p_{EI}^2\end{aligned} \tag{34}$$

$$\begin{aligned}\frac{dp_{EI}}{dt} &= \big(r_I\big(r_E-\bar{r}_E\big)S^I + c_{EI}\sigma^2 S^{EI}_\eta\big)p_0^{EI} + S_F^{EI}\epsilon p_{EI} + S_B^{EI}\epsilon p_0^{EI}p_{IE}\\ &\quad + S_C^{EIE}\epsilon^2 p_0^{EI}p_{IE}\left(\kappa p_{AA} + (N_E-\kappa)\,p_{AB}\right) + S_C^{EII}\epsilon^2 p_0^{EI}N_I p_{EI}p_{II}.\end{aligned} \tag{35}$$

Finally, we recall that $\epsilon = (\kappa p_0)^{-1}$ and $N_I = \gamma\kappa$, revealing that the dynamics above stop at $\mathcal{O}(\epsilon)$:

$$\frac{dp_{AA}}{dt} = \left(r_E^2 S + c_{AA}\sigma^2 S_\eta\right) p_0 + \epsilon\left[S_F p_{AA} + S_B\left(q_{AA} + p_0 p_{AA}\right) + S_C\left(p_{AA}^2 + (M-1)\,p_{AB}^2\right) + S_C^I \gamma p_{EI}^2\right] \tag{36}$$

$$\frac{dp_{AB}}{dt} = \left(r_E^2 S + c_{AB}\sigma^2 S_\eta\right) p_0 + \epsilon\big[S_F p_{AB} + S_B(q_{AB} + p_0 p_{AB}) + S_C\left(2p_{AA}p_{AB} + (M-2)\,p_{AB}^2\right) + S_C^I \gamma p_{EI}^2\big] \tag{37}$$

$$\frac{dq_{AA}}{dt} = \left(r_E^2 S + c_{AA}\sigma^2 S_\eta\right) q_0 + \epsilon\big[S_F q_{AA} + S_B(1-p_0)(q_{AA} + p_0 p_{AA}) + S_C \frac{1}{p_0} q_0\left(p_{AA}^2 + (M-1)p_{AB}^2\right) + S_C^I \frac{\gamma}{p_0} q_0 p_{EI}^2\big] \tag{38}$$

$$\frac{dq_{AA}}{dt} = \left(r_E^2 S + c_{AA}\sigma^2 S_\eta\right) q_0 + \epsilon\big[S_F q_{AA} + S_B(1-p_0)(q_{AA} + p_0 p_{AA}) + S_C \frac{1}{p_0} q_0\left(p_{AA}^2 + (M-1)p_{AB}^2\right) + S_C^I \frac{\gamma}{p_0} q_0 p_{EI}^2\big] \tag{39}$$

$$\frac{dq_{AB}}{dt} = \left(r_E^2 S + c_{AB}\sigma^2 S_\eta\right) q_0 + \epsilon\big[S_F q_{AB} + S_B\left(1-p_0\right)\left(q_{AB} + p_0 p_{AB}\right) + S_C \frac{1}{p_0} q_0\left(2p_{AA}p_{AB} + (M-2)\,p_{AB}^2\right) + S_C^I \frac{\gamma}{p_0} q_0 p_{EI}^2\big] \tag{40}$$

$$\frac{dp_{EI}}{dt} = \left(r_I\left(r_E - \bar{r}_E\right)S^I + c_{EI}\sigma^2 S_\eta^{EI}\right) p_0^{EI} + \epsilon\big[S_F^{EI} p_{EI} + S_B^{EI} p_0^{EI} p_{IE} + S_C^{EIE}\frac{p_0^{EI}}{p_0} p_{IE}\left(p_{AA} + (M-1)\,p_{AB}\right) + S_C^{EII}\frac{p_0^{EI}}{p_0}\gamma p_{EI} p_{II}\big]. \tag{41}$$

5.4 Firing rate dynamics

Here we have written the dynamics in terms of the average firing rates r_E, r_I and STDP-weighted spiking covariances as if those were parameters. As the mean excitatory and inhibitory weights change, so will neurons' firing rates. We now supplement the dynamics of the connectivity by examining the evolution of the population-averaged firing rates r_α with $\alpha \in \{E, I\}$. The quasi-stationary firing rates obey:

$$r_\alpha(t) = f_\alpha(\mu_\alpha(t), \sigma^2) \tag{42}$$

where f_α is the rate-current function of an EIF neuron belonging to population α and

$$\begin{aligned}
\mu_E &= \mu_{\text{ext},E} + \epsilon N_E p \tau_E r_E + \epsilon N_I p_{EI} \tau_I r_I \\
&= \mu_{\text{ext},E} + \epsilon \left(\kappa p_{AA} + (N_E - \kappa) p_{AB}\right) \tau_E r_E + \epsilon N_I p_{EI} \tau_I r_I \\
&= \mu_{\text{ext},E} + \frac{1}{p_0} \left(p_{AA} + (M-1)\, p_{AB}\right) \tau_E r_E + \frac{\gamma}{p_0} p_{EI} \tau_I r_I \\
\mu_I &= \mu_{\text{ext},I} + \epsilon N_E p_{IE} \tau_E r_E + \epsilon N_I p_{II} \tau_I r_I \\
&= \mu_{\text{ext},I} + \frac{M}{p_0} p_{IE} \tau_E r_E + \frac{\gamma}{p_0} p_{II} \tau_I r_I
\end{aligned} \tag{43}$$

is the average external input to one of those neurons and we assume that a sufficient combination of low firing rates and weak/slow synapses keeps recurrent connectivity from contributing significantly to the effective variance of inputs to a neuron. τ_E and τ_I are the integrals of excitatory and inhibitory synaptic kernels (these are described by single exponentials, so the integral is their decay time constant).

The dynamics of the quasi-stationary firing rates is then given by:

$$\frac{dr_\alpha}{dt} = \frac{df_\alpha}{d\mu_\alpha} \frac{d\mu_\alpha}{dt}. \tag{44}$$

Recalling that $\left.\frac{df_\alpha}{d\mu}\right|_{\mu_\alpha} = \int_0^\infty A_\alpha(t) dt$, where $A_\alpha(t)$ is the average linear response of neurons of type α, we define

$$S_A^\alpha \equiv \int_0^\infty A_\alpha(t) dt. \tag{45}$$

Assuming that $\mu_{\text{ext},\alpha}$ is constant in time, we obtain:

$$\frac{dr_E}{dt} = S_A^E \left(\frac{\tau_E}{p_0} \left(\left(p_{AA} + (M-1)\, p_{AB}\right) \frac{dr_E}{dt} + \left(\frac{dp_{AA}}{dt} + (M-1) \frac{dp_{AB}}{dt} \right) r_E \right) + \frac{\gamma \tau_I}{p_0} \left(\frac{dp_{EI}}{dt} r_I + p_{EI} \frac{dr_I}{dt} \right) \right) \tag{46}$$

and since the excitatory $\rightarrow$ inhibitory and inhibitory $\rightarrow$ inhibitory weights are not plastic, r_I tracks r_E:

$$\begin{aligned}
\frac{dr_I}{dt} &= S_A^I \left(\frac{M}{p_0} p_{IE} \tau_E \frac{dr_E}{dt} + \frac{\gamma}{p_0} p_{II} \tau_I \frac{dr_I}{dt} \right) \\
&= \left(\frac{S_A^I \frac{M}{p_0} p_{IE} \tau_E}{1 - \frac{\gamma}{p_0} p_{II} \tau_I} \right) \frac{dr_E}{dt}.
\end{aligned} \tag{47}$$

Inserting Eq. (47) into Eq. (46) then yields

$$\frac{dr_E}{dt} = \frac{S_A^E \left(\frac{\tau_E}{p_0} \left(\frac{dp_{AA}}{dt} + (M-1) \frac{dp_{AB}}{dt} \right) r_E + \frac{\gamma \tau_I}{p_0} \frac{dp_{EI}}{dt} r_I \right)}{\left(1 - S_A^E \frac{\tau_E}{p_0} \left(p_{AA} + (M-1)\, p_{AB}\right) - S_A^E \frac{\gamma \tau_I}{p_0} p_{EI} \left(\frac{S_A^I \frac{M}{p_0} p_{IE} \tau_E}{1 - \frac{\gamma}{p_0} p_{II} \tau_I} \right) \right)}. \tag{48}$$

5.5 Linear stability of firing rates

The iSTDP rule imposes a form of rate homeostasis on the dynamics, keeping r_E within $\mathcal{O}(\epsilon)$ of $\bar{r}_E$. Indeed, this was one major motivation for its theoretical proposal (Sprekeler & Vogels et al, 2011). We now check how this affects the dynamics of the weights. If there is a balance between potentiation and depression in the eSTDP rule $L(s)$ so that $S \sim \mathcal{O}(\epsilon)$, then the dynamics of mean excitatory weights have an $\mathcal{O}(1/\epsilon)$ timescale. There is a different condition for balance between potentiation and depression of inhibitory $\rightarrow$ excitatory synapses. This balance occurs when the excitatory rate is close to $\bar{r}_E$, requiring $(r_E - \bar{r}_E) \sim \mathcal{O}(\epsilon)$. If the eSTDP rule is balanced but $(r_E - \bar{r}_E) \sim \mathcal{O}(1)$ then the leading order dynamics of the firing rates and p_{EI} become $\mathcal{O}(1)$ and obey Eq. (4):

$$\frac{dr_E}{dt} = \underbrace{\left(\frac{S_A^E \frac{\gamma \tau_I}{p_0}}{1 - S_A^E \frac{\tau_E}{p_0} \left(p_{AA} + (M-1)\, p_{AB}\right) - S_A^E \frac{\gamma \tau_I}{p_0} p_{EI} \left(\frac{S_A^I \frac{M}{p_0} p_{IE} \tau_E}{1 - \frac{\gamma}{p_0} p_{II} \tau_I} \right)} \right)}_{X(p_{EI})} r_I \frac{dp_{EI}}{dt} \tag{49}$$

$$\frac{dr_I}{dt} = \underbrace{\left(\frac{S_A^I \frac{M}{p_0} p_{IE} \tau_E}{1 - \frac{\gamma}{p_0} p_{II} \tau_I} \right)}_{Y} \frac{dr_E}{dt} \tag{50}$$

with fixed points (p_{EI}^*, r_I^*, r_E^*) obeying:

$$\begin{aligned} 0 &= \left(r_I^* \left(r_E^* - \bar{r}_E \right) S^I + c_{EI} \sigma^2 S_\eta^{EI} \right) \\ 0 &= X(p_{EI}^*) \cdot r_I^* \cdot \left(r_I^* \left(r_E^* - \bar{r}_E \right) S^I + c_{EI} \sigma^2 S_\eta^{EI} \right) \\ 0 &= Y \cdot X(p_{EI}^*) \cdot r_I^* \cdot \left(r_I^* \left(r_E^* - \bar{r}_E \right) S^I + c_{EI} \sigma^2 S_\eta^{EI} \right). \end{aligned} \tag{51}$$

In order for the first condition to hold ($dp_{EI}/dt = 0$), the fixed point rates must lie on the hyperbola given by

$$r_E^* = -\frac{c_{EI} \sigma^2 S_\eta^{EI}}{S^I} \left(\frac{1}{r_I^*} \right) + \bar{r}_E. \tag{52}$$

This also satisfies $dr_E/dt = 0$ and $dr_I/dt = 0$. If $c_{EI} = 0$, this reduces to $r_E^* = \bar{r}_E$.

We next examine the linear stability of this solution. The Jacobian for Eqs. (4), (49)–(50) is:

$$\begin{pmatrix} 0 & r_I S^I p_0^{EI} & (r_E - \bar{r}_E) S^I p_0^{EI} \\ r_I \left(r_I \left(r_E - \bar{r}_E \right) S^I + c_{EI} \sigma^2 S_\eta^{EI} \right) p_0^{EI} \frac{\partial X}{\partial p_{EI}} & X r_I^2 S^I p_0^{EI} & 2 X p_0^{EI} (r_E - \bar{r}_E) S^I r_I + X c_{EI} \sigma^2 S_\eta^{EI} \\ Y r_I \left(r_I \left(r_E - \bar{r}_E \right) S^I + c_{EI} \sigma^2 S_\eta^{EI} \right) p_0^{EI} \frac{\partial X}{\partial p_{EI}} & Y X r_I^2 S^I p_0^{EI} & 2 Y X p_0^{EI} (r_E - \bar{r}_E) S^I r_I + Y X c_{EI} \sigma^2 S_\eta^{EI} \end{pmatrix}$$

where

$$\frac{\partial X}{\partial p_{EI}} = \frac{\left(S_A^E \frac{\gamma \tau_I}{p_0}\right)^2 \left(\frac{S_A^I \frac{M}{p_0} p_{IE} \tau_E}{1-\frac{\gamma}{p_0} p_{II} \tau_I}\right)}{\left(1 - S_A^E \frac{\tau_E}{p_0}\left(p_{AA} + (M-1)\, p_{AB}\right) - S_A^E \frac{\gamma \tau_I}{p_0} p_{EI} \left(\frac{S_A^I \frac{M}{p_0} p_{IE} \tau_E}{1-\frac{\gamma}{p_0} p_{II} \tau_I}\right)\right)^2} \tag{53}$$

The eigenvalues of the Jacobian, evaluated at $p_{EI}^*, r_E^* = \bar{r}_E, r_I^*$ with $c_{EI} = 0$, are:

$$\begin{aligned} \lambda_1 &= \lambda_2 = 0, \\ \lambda_3 &= \frac{\gamma \left(r_I^*\right)^2 S_A^E S^I \tau_I}{1 - \frac{S_A^E \tau_E (p_{AA} + (M-1) p_{AB})}{p_0} - \frac{\gamma M p_{EI} p_{IE} S_A^E S_A^I \tau_E \tau_I}{p_0^2 - \gamma p_0 p_{II} \tau_I}} \end{aligned} \tag{54}$$

Below, we plot these eigenvalues (with $c_{EI} = 0$) as a function of the total excitation $p_{AA} + (M-1)p_{AB}$ with $c_{EI} = 0$ so that $r_E^* = \bar{r}_E$. For each $p_{AA} + (M-1)p_{AB}$, we use bisection to find $p_{EI}^* \in [0, \mathrm{W}_{\max}^I]$ that minimizes $|(r_E - \bar{r}_E)|$ (for the particular cellular and network parameters used). Fortunately, the inhibition is strong enough to achieve $r_E = \bar{r}_E$ - it would be possible for this not to be the case, for example with weak $\mathrm{W}_{\max}^I$.

5.6 Final dynamics of network structure: mind your p's and q's

The above analysis of unbalanced iSTDP reveals that there is a $\mathcal{O}(\epsilon)$ neighborhood around $p_{EI}^*, \bar{r}_E, r_I^*$ which is attracting along those dimensions, so that $r_E = \bar{r}_E + \mathcal{O}(\epsilon), r_I = r_I^* + \mathcal{O}(\epsilon), p_{EI} = p_{EI}^* + \mathcal{O}(\epsilon)$. (If $c_{EI} \neq 0$ then $\lambda_2 \neq 0$ and the dynamics could be different, a potential subject for future study.) Inserting these yields the following equations, up to $\mathcal{O}(\epsilon)$ and for balanced eSTDP (so $S \sim \mathcal{O}(\epsilon)$):

$$\frac{dp_{AA}}{dt} = \left(\bar{r}_E^2 S + c_{AA}\sigma^2 S_\eta\right) p_0 + \epsilon \left[S_F p_{AA} + S_B \left(q_{AA} + p_0 p_{AA}\right) + S_C \left(p_{AA}^2 + (M-1)\, p_{AB}^2\right) + S_C^I \gamma (p_{EI}^*)^2\right] \tag{55}$$

$$\begin{aligned} \frac{dp_{AB}}{dt} = \left(\bar{r}_E^2 S + c_{AB}\sigma^2 S_\eta\right) p_0 + \epsilon \big[& S_F p_{AB} + S_B (q_{AB} + p_0 p_{AB}) + S_C \left(2 p_{AA} p_{AB} + (M-2)\, p_{AB}^2\right) \\ & + S_C^I \gamma (p_{EI}^*)^2 \big] \end{aligned} \tag{56}$$

$$\begin{aligned} \frac{dq_{AA}}{dt} = \left(\bar{r}_E^2 S + c_{AA}\sigma^2 S_\eta\right) q_0 + \epsilon \big[& S_F q_{AA} + S_B (1-p_0)(q_{AA} + p_0 p_{AA}) \\ & + S_C \frac{1}{p_0} q_0 \left(p_{AA}^2 + (M-1) p_{AB}^2\right) + S_C^I \frac{\gamma}{p_0} q_0 (p_{EI}^*)^2 \big] \end{aligned} \tag{57}$$

$$\begin{aligned} \frac{dq_{AB}}{dt} = \left(\bar{r}_E^2 S + c_{AB}\sigma^2 S_\eta\right) q_0 + \epsilon \big[& S_F q_{AB} + S_B \left(1-p_0\right)\left(q_{AB} + p_0 p_{AB}\right) \\ & + S_C \frac{1}{p_0} q_0 \left(2 p_{AA} p_{AB} + (M-2)\, p_{AB}^2\right) + S_C^I \frac{\gamma}{p_0} q_0 (p_{EI}^*)^2 \big]. \end{aligned} \tag{58}$$

Note that the location of $(\bar{r}_E, r_I^*, p_{EI}^*)$ depends on the net excitation, $p_{AA} + (M-1)p_{AB}$, and so will evolve on the slow timescale of the balanced eSTDP. We compute the nullclines of these

equations in asymmetric networks by bisection. For example, for each p_{AA} we find the p_{AB} for which the homeostatic p^*_{EI} associated with (p_{AA}, p_{AB}) gives $dp_{AA}/dt = 0$.

5.7 Temporally symmetric eSTDP

When the timescales of potentiation and depression in the excitatory STDP rule are similar, $\tau_+ \sim \tau_- + \mathcal{O}(\epsilon)$, then the dynamics of the network structure simplify considerably. Since the correlations from common inputs (both from excitatory and inhibitory neurons) are temporally symmetric around 0 lag, this makes $S_C, S_C^I, S_C^{EIE}, S_C^{EII} \sim \mathcal{O}(\epsilon)$. The dynamics then reduce to:

$$\frac{dp_\alpha}{dt} = \left(\bar{r}_E^2 S + c_\alpha \sigma^2 S_\eta\right) p_0 + \epsilon \left[S_F p_\alpha + S_B \left(q_\alpha + p_0 p_\alpha\right)\right] \tag{59}$$

$$\frac{dq_\alpha}{dt} = \left(\bar{r}_E^2 S + c_\alpha \sigma^2 S_\eta\right) q_0 + \epsilon \big[S_F q_\alpha + S_B (1 - p_0)(q_\alpha + p_0 p_\alpha)\big] \tag{60}$$

for $\alpha = AA$ or AB.

5.8 Separable dynamics of assembly formation and segregation

The dynamics of the network structure simplify if we take a linear transformation of our p and q variables:

$$\begin{aligned} p &= \frac{M p_{AA} + M(M-1) p_{AB}}{M^2} \\ q &= \frac{M q_{AA} + M(M-1) q_{AB}}{M^2} \\ p_\Delta &= p_{AA} - p_{AB} \\ q_\Delta &= q_{AA} - q_{AB}. \end{aligned} \tag{61}$$

The first two, p, q, measure the total mean synaptic weight and the mean weight of reciprocal connections overall in the network. The second two measure the formation of structure. The dynamics of these transformed variables are:

$$\frac{dp}{dt} = \left(\bar{r}_E^2 S + c_{EE} \sigma^2 S_\eta\right) p_0 + \epsilon \left[S_F p + S_B \left(q + p_0 p\right) + S_C M p^2 + S_C^I \gamma (p^*_{EI})^2\right] \tag{62}$$

$$\frac{dq}{dt} = \left(\bar{r}_E^2 S + c_{EE} \sigma^2 S_\eta\right) q_0 + \epsilon \left[S_F q + S_B \left(1 - p_0\right)\left(q + p_0 p\right) + S_C M \frac{q_0}{p_0} p^2 + S_C^I \gamma \frac{q_0}{p_0} (p^*_{EI})^2\right] \tag{63}$$

$$\frac{dp_\Delta}{dt} = c_\Delta \sigma^2 S_\eta p_0 + \epsilon \left[S_F p_\Delta + S_B (q_\Delta + p_0 p_\Delta) + S_C p_\Delta^2\right] \tag{64}$$

$$\frac{dq_\Delta}{dt} = c_\Delta \sigma^2 S_\eta q_0 + \epsilon \left[S_F q_\Delta + S_B \left(1 - p_0\right)\left(q_\Delta + p_0 p_\Delta\right) + S_C \frac{q_0}{p_0} p_\Delta^2\right] \tag{65}$$

where c_{EE} is defined, analogously to p, as the average correlation of external inputs and $c_\Delta = c_{AA} - c_{AB}$. Here we see that the spontaneous dynamics of overall potentation/depression (p, q) are separable from the dynamics of structure formation (p_Δ, q_Δ).

The nullclines are given by solving each equation for the steady state, and are:

$$
\begin{aligned}
p^* &= \frac{-\epsilon(S_F + p_0 S_B) \pm \sqrt{(\epsilon(S_F + p_0 S_B))^2 - 4\epsilon S_C M \left(\left(\bar{r}_E^2 S + c_{EE}\sigma^2 S_\eta\right) p_0 + \epsilon(S_B q^* + S_C^I \gamma (p_{EI}^*)^2)\right)}}{2\epsilon S_C M} \\
q^* &= -\frac{\left(\bar{r}_E^2 S + c_{EE}\sigma^2 S_\eta\right) q_0 + \epsilon\left(S_B(1-p_0)p_0 p^* + S_C M \frac{q_0}{p_0}(p^*)^2 + S_C^I \gamma \frac{q_0}{p_0}(p_{EI}^*)^2\right)}{\epsilon(S_F + (1-p_0)S_B)}
\end{aligned}
\tag{66}
$$

$$
\begin{aligned}
p_\Delta^* &= \frac{-\epsilon(S_F + p_0 S_B) \pm \sqrt{\epsilon^2 (S_F + p_0 S_B)^2 - 4\epsilon S_C \left(c_\Delta \sigma^2 S_\eta p_0 + \epsilon S_B q_\Delta^*\right)}}{2\epsilon S_C} \\
q_\Delta^* &= -\frac{c_\Delta \sigma^2 S_\eta q_0 + \epsilon\left(S_B(1-p_0)p_0 p_\Delta^* + S_C \frac{q_0}{p_0}(p_\Delta^*)^2\right)}{\epsilon\left(S_F + S_B(1-p_0)\right)}
\end{aligned}
\tag{67}
$$

In the spontaneous case ($c_{EE} = c_\Delta = 0$) and defining $S = -\delta\epsilon$ these simplify to:

$$
\begin{aligned}
p^* &= \frac{-(S_F + p_0 S_B) \pm \sqrt{((S_F + p_0 S_B))^2 - 4 S_C M \left(-\bar{r}_E^2 \delta p_0 + S_B q^* + S_C^I \gamma (p_{EI}^*)^2\right)}}{2 S_C M} \\
q^* &= -\frac{-\bar{r}_E^2 \delta q_0 + S_B(1-p_0)p_0 p^* + S_C M \frac{q_0}{p_0}(p^*)^2 + S_C^I \gamma \frac{q_0}{p_0}(p_{EI}^*)^2}{S_F + (1-p_0)S_B}
\end{aligned}
\tag{68}
$$

$$
\begin{aligned}
p_\Delta^* &= \frac{-(S_F + p_0 S_B) \pm \sqrt{(S_F + p_0 S_B)^2 - 4 S_C \left(S_B q_\Delta^*\right)}}{2 S_C} \\
q_\Delta^* &= -\frac{S_B(1-p_0)p_0 p_\Delta + S_C \frac{q_0}{p_0}(p_\Delta^*)^2}{S_F + (1-p_0)S_B}.
\end{aligned}
\tag{69}
$$

5.9 Fisher Information

We consider the linear Fisher information of an assembly's activity about a stimulus θ:

$$
FI_A \approx \left(\frac{dn_A}{d\theta}\right)^2 C_{AA}^{-1} = \left(\kappa T \frac{dr_A}{d\theta}\right)^2 C_{AA}^{-1} \tag{70}
$$

where C_{AA} is the variance of an assembly's summed spike count in a window of $T = 100$ ms. We compute it, using the length one approximation of spike train covariances (23) as [101]:

$$
\begin{aligned}
C_{AA} &= \sum_{i,j \in A} \left(\int_{-T}^{T} (T - |s|)\mathbf{C}_{ij}(s))\, ds\right) \\
&= \int_{-T}^{T} (T - |s|) \Big[\kappa C_E^0(s) + \kappa(\kappa - 1)\Big(c_{AA}\sigma^2 \left(A_E(t) * A_E(-t)\right) + p_{AA}\left(K_{EE} * C_E^0(s)\right) \\
&\quad + \left(q_{AA} + p_0 p_{AA}\right) C_E^0(s) * K_{EE}(-t) + \left(\kappa p_{AA}^2 + (N_E - \kappa) p_{AB}^2\right)\left(K_{EE}(t) * C_E^0(s) * K_{EE}(-t)\right) \\
&\quad + N_I \left(p_{EI}^*\right)^2 \left(K_{EI}(t) * C_I^0(s) * K_{EI}(-t)\right)\Big)\Big] ds
\end{aligned}
\tag{71}
$$

In order to calculate the stimulus-response gain $\frac{dr_A}{d\theta}$, we consider all sources of input to neurons in assembly A:

$$\mu_{\text{tot,A}} = \mu_{\text{ext}} + \mu_\theta + \frac{\tau_E}{p_0} p_{AA} r_A + \frac{\tau_E}{p_0}(M-1) p_{AB} r_B + \frac{\gamma \tau_I}{p_0} p^*_{EI} r_I \tag{72}$$

(using $\epsilon = (\kappa p_0)^{-1}$), where r_A and r_B are the rates of excitatory neurons in the stimulated (r_A) or non-stimulated (r_B) assemblies. The total inputs to excitatory neurons in non-stimulated assemblies and inhibitory neurons are:

$$\begin{aligned} \mu_{\text{tot,B}} &= \mu_{\text{ext}} + \frac{\tau_E}{p_0}\left(p_{AA} r_B + p_{AB} r_A + (M-2) p_{AB} r_B\right) + \frac{\gamma \tau_I}{p_0} p^*_{EI} r_I \\ \mu_{\text{tot,I}} &= \mu_{\text{ext}} + \frac{\tau_E}{p_0} p_{IE}\left(r_A + (M-1) r_B\right) + \frac{\gamma \tau_I}{p_0} p_{II} r_I . \end{aligned} \tag{73}$$

Applying the chain rule gives:

$$\begin{aligned} \frac{dr_A}{d\theta} &= \frac{dr_A}{d\mu_{\text{tot,A}}} \frac{d\mu_{\text{tot,A}}}{d\theta} \\ &= \left(\frac{dr_A}{d\mu_{\text{tot,A}}}\right)\left(\frac{d\mu_\theta}{d\theta} + \frac{\tau_E}{p_0}\left(p_{AA}\frac{dr_A}{d\theta} + (M-1) p_{AB}\frac{dr_B}{d\theta}\right) + \frac{\gamma\tau_I}{p_0} p^*_{EI}\frac{dr_I}{d\theta}\right) \\ &= \frac{\frac{dr_A}{d\mu_{\text{tot,A}}}}{\left(1 - \frac{dr_A}{d\mu_{\text{tot,A}}}\frac{\tau_E}{p_0} p_{AA}\right)}\left(\frac{d\mu_\theta}{d\theta} + \frac{\tau_E}{p_0}(M-1) p_{AB}\frac{dr_B}{d\theta} + \frac{\gamma\tau_I}{p_0} p^*_{EI}\frac{dr_I}{d\theta}\right) \end{aligned} \tag{74}$$

$$\begin{aligned} \frac{dr_B}{d\theta} &= \frac{dr_B}{d\mu_{\text{tot,B}}} \frac{d\mu_{\text{tot,B}}}{d\theta} \\ &= \frac{dr_B}{d\mu_{\text{tot,B}}}\left(\frac{\tau_E}{p_0}\left(p_{AA}\frac{dr_B}{d\theta} + p_{AB}\frac{dr_A}{d\theta} + (M-2) p_{AB}\frac{dr_B}{d\theta}\right) + \frac{\gamma\tau_I}{p_0} p^*_{EI}\frac{dr_I}{d\theta}\right) \\ &= \frac{\frac{dr_B}{d\mu_{\text{tot,B}}}}{\left(1 - \frac{dr_B}{d\mu_{\text{tot,B}}}\frac{\tau_E}{p_0}\left(p_{AA} + (M-2) p_{AB}\right)\right)}\left(\frac{\tau_E}{p_0} p_{AB}\frac{dr_A}{d\theta} + \frac{\gamma\tau_I}{p_0} p^*_{EI}\frac{dr_I}{d\theta}\right) \end{aligned} \tag{75}$$

$$\begin{aligned} \frac{dr_I}{d\theta} &= \frac{dr_I}{d\mu_{\text{tot,I}}} \frac{d\mu_{\text{tot,I}}}{d\theta} \\ &= \frac{dr_I}{d\mu_{\text{tot,I}}}\left(\frac{\tau_E}{p_0} p^*_{EI}\left(\frac{dr_A}{d\theta} + (M-1)\frac{dr_B}{d\theta}\right) + \frac{\gamma\tau_I}{p_0} p_{II}\frac{dr_I}{d\theta}\right) \\ &= \frac{\frac{dr_I}{d\mu_{\text{tot,I}}}}{1 - \frac{dr_I}{d\mu_{\text{tot,I}}}\frac{\gamma\tau_I}{p_0} p_{II}}\left(\frac{\tau_E}{p_0} p^*_{EI}\left(\frac{dr_A}{d\theta} + (M-1)\frac{dr_B}{d\theta}\right)\right). \end{aligned} \tag{76}$$

Inserting $dr_I/d\theta$ into $dr_B d\theta$:

$$\frac{dr_B}{d\theta} = \frac{\frac{dr_B}{d\mu_{\text{tot,B}}}}{\left(1 - \frac{dr_B}{d\mu_{\text{tot,B}}}\frac{\tau_E}{p_0}\left(p_{AA} + (M-2)p_{AB}\right)\right)}\left(\frac{\tau_E}{p_0}p_{AB}\frac{dr_A}{d\theta} + \frac{\gamma\tau_I}{p_0}p^*_{EI}\frac{\frac{dr_I}{d\mu_{\text{tot,I}}}}{1 - \frac{dr_I}{d\mu_{\text{tot,I}}}\frac{\gamma\tau_I}{p_0}p_{II}}\left(\frac{\tau_E}{p_0}p^*_{EI}\left(\frac{dr_A}{d\theta} + (M-1)\right.\right.\right.$$

$$\frac{dr_B}{d\theta} = \frac{dr_A}{d\theta}\underbrace{\frac{\frac{dr_B}{d\mu_{\text{tot,B}}}\left(\frac{\tau_E}{p_0}p_{AB} + \frac{\gamma\tau_I}{p_0}p^*_{EI}\left(\frac{\frac{dr_I}{d\mu_{\text{tot,I}}}}{1-\frac{dr_I}{d\mu_{\text{tot,I}}}\frac{\gamma\tau_I}{p_0}p_{II}}\right)\frac{\tau_E}{p_0}p^*_{EI}\right)}{\left(1 - \frac{dr_B}{d\mu_{\text{tot,B}}}\frac{\tau_E}{p_0}\left(p_{AA} + (M-2)p_{AB}\right) - \frac{\gamma\tau_I}{p_0}p^*_{EI}\frac{\frac{dr_I}{d\mu_{\text{tot,I}}}}{1-\frac{dr_I}{d\mu_{\text{tot,I}}}\frac{\gamma\tau_I}{p_0}p_{II}}\frac{\tau_E}{p_0}p^*_{EI}(M-1)\right)}}_{\equiv X} \tag{77}$$

and inserting $dr_B/d\theta$ into $dr_I/d\theta$:

$$\frac{dr_I}{d\theta} = \frac{dr_A}{d\theta}\underbrace{\frac{\frac{dr_I}{d\mu_{\text{tot,I}}}}{1 - \frac{dr_I}{d\mu_{\text{tot,I}}}\frac{\gamma\tau_I}{p_0}p_{II}}\left(\frac{\tau_E}{p_0}p^*_{EI}\left(1 + (M-1)X\right)\right)}_{\equiv Y} \tag{78}$$

which yields:

$$\begin{aligned}\frac{dr_A}{d\theta} &= \frac{\frac{dr_A}{d\mu_{\text{tot,A}}}}{\left(1 - \frac{dr_A}{d\mu_{\text{tot,A}}}\frac{\tau_E}{p_0}p_{AA}\right)}\left(\frac{d\mu_\theta}{d\theta} + \frac{\tau_E}{p_0}(M-1)p_{AB}X\frac{dr_A}{d\theta} + \frac{\gamma\tau_I}{p_0}p^*_{EI}Y\frac{dr_A}{d\theta}\right)\\ \frac{dr_A}{d\theta} &= \frac{\frac{dr_A}{d\mu_{\text{tot,A}}}\frac{d\mu_\theta}{d\theta}}{\left(1 - \frac{dr_A}{d\mu_{\text{tot,A}}}\frac{\tau_E}{p_0}p_{AA} - \frac{\tau_E}{p_0}(M-1)p_{AB}X - \frac{\gamma\tau_I}{p_0}p^*_{EI}Y\right)}.\end{aligned} \tag{79}$$

6 Acknowledgments

We thank Ken Miller, Bard Ermentrout, Jon Rubin, Anne-Marie Oswald and Thanos Tzounopoulos for their valuable comments.

References

1. Hebb DO (1949) The organization of behavior: a neuropsychological theory. Mahwah, N.J.: L. Erlbaum Associates.

2. Markram H, Gerstner W, Sjöström PJ (2011) A history of spike-timing-dependent plasticity. Frontiers in Synaptic Neuroscience 3: 4.

3. Harris KD, Mrsic-Flogel TD (2013) Cortical connectivity and sensory coding. Nature 503: 51–58.

4. Buzsáki G (2010) Neural syntax: cell assemblies, synapsembles, and readers. Neuron 68: 362–385.

5. Neves G, Cooke SF, Bliss TV (2008) Synaptic plasticity, memory and the hippocampus: a neural network approach to causality. Nature Reviews Neuroscience 9: 65–75.

6. Ko H, Hofer SB, Pichler B, Buchanan KA, Sjstrm PJ, et al. (2011) Functional specificity of local synaptic connections in neocortical networks. Nature 473: 87–91.

7. Cossell L, Iacaruso MF, Muir DR, Houlton R, Sader EN, et al. (2015) Functional organization of excitatory synaptic strength in primary visual cortex. Nature 518: 399–403.

8. Lee WCA, Bonin V, Reed M, Graham BJ, Hood G, et al. (2016) Anatomy and function of an excitatory network in the visual cortex. Nature 532: 370–374.

9. Yoshimura Y, Dantzker JLM, Callaway EM (2005) Excitatory cortical neurons form fine-scale functional networks. Nature 433: 868–873.

10. Perin R, Berger TK, Markram H (2011) A synaptic organizing principle for cortical neuronal groups. Proceedings of the National Academy of Sciences 108: 5419–5424.

11. Ko H, Cossell L, Baragli C, Antolik J, Clopath C, et al. (2013) The emergence of functional microcircuits in visual cortex. Nature 496: 96–100.

12. Ko H, Mrsic-Flogel TD, Hofer SB (2014) Emergence of Feature-Specific Connectivity in Cortical Microcircuits in the Absence of Visual Experience. The Journal of Neuroscience 34: 9812–9816.

13. Markram H, Gerstner W, Sjöström PJ (2012) Spike-timing-dependent plasticity: a comprehensive overview. Frontiers in Synaptic Neuroscience : 2.

14. Masuda N, Kori H (2007) Formation of feedforward networks and frequency synchrony by spike-timing-dependent plasticity. Journal of computational neuroscience 22: 327–345.

15. Takahashi YK, Kori H, Masuda N (2009) Self-organization of feed-forward structure and entrainment in excitatory neural networks with spike-timing-dependent plasticity. Physical Review E 79: 051904.

16. Tannenbaum NR, Burak Y (2016) Shaping neural circuits by high order synaptic interactions. arXiv preprint arXiv:160503005 .

17. Gerstner W, Kempter R, van Hemmen JL, Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. Nature 383: 76–78.

18. Song S, Abbott LF (2001) Cortical development and remapping through spike timing-dependent plasticity. Neuron 32: 339–350.

19. Gerstner W, Ritz R, van Hemmen JL (1993) Why spikes? Hebbian learning and retrieval of time-resolved excitation patterns. Biological Cybernetics 69: 503–515.

20. Fiete IR, Senn W, Wang CZH, Hahnloser RHR (2010) Spike-Time-Dependent Plasticity and Heterosynaptic Competition Organize Networks to Produce Long Scale-Free Sequences of Neural Activity. Neuron 65: 563–576.

21. Mongillo G, Curti E, Romani S, Amit DJ (2005) Learning in realistic networks of spiking neurons and spike-driven plastic synapses. The European Journal of Neuroscience 21: 3143–3160.

22. Litwin-Kumar A, Doiron B (2014) Formation and maintenance of neuronal assemblies through synaptic plasticity. Nature Communications 5.

23. Zenke F, Agnes EJ, Gerstner W (2015) Diverse synaptic plasticity mechanisms orchestrated to form and retrieve memories in spiking neural networks. Nature Communications 6.

24. Pfister JP, Gerstner W (2006) Triplets of Spikes in a Model of Spike Timing-Dependent Plasticity. The Journal of Neuroscience 26: 9673–9682.

25. Clopath C, Gerstner W (2010) Voltage and Spike Timing Interact in STDP A Unified Model. Frontiers in Synaptic Neuroscience 2.

26. Graupner M, Brunel N (2012) Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location. Proceedings of the National Academy of Sciences 109: 3991–3996.

27. Sjöström PJ, Turrigiano GG, Nelson SB (2001) Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. Neuron 32: 1149–1164.

28. Gilson M, Burkitt AN, Grayden DB, Thomas DA, Hemmen JL (2009) Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks IV: Structuring synaptic pathways among recurrent connections. Biological Cybernetics 101: 427–444.

29. Clopath, Claudia, Bsing, Lars, Vasilaki, Eleni, Gerstner, Wulfram (2010) Connectivity reflects coding: a model of voltage-based STDP with homeostasis. Nat Neurosci 13: 344–352.

30. Bair W, Zohary E, Newsome WT (2001) Correlated firing in macaque visual area MT: time scales and relationship to behavior. J Neurosci 21: 1676–1697.

31. Kohn A, Smith MA (2005) Stimulus Dependence of Neuronal Correlation in Primary Visual Cortex of the Macaque. The Journal of Neuroscience 25: 3661–3673.

32. Rothschild G, Nelken I, Mizrahi A (2010) Functional organization and population dynamics in the mouse primary auditory cortex. Nature Neuroscience 13: 353–360.

33. Börgers C, Franzesi GT, LeBeau FE, Boyden ES, Kopell NJ (2012) Minimal size of cell assemblies coordinated by gamma oscillations. PLoS Comput Biol 8: e1002362.

34. Jia X, Tanabe S, Kohn A (2013) Gamma and the coordination of spiking activity in early visual cortex. Neuron 77: 762–774.

35. Salkoff DB, Zagha E, Yüzgeç Ö, McCormick DA (2015) Synaptic mechanisms of tight spike synchrony at gamma frequency in cerebral cortex. The Journal of Neuroscience 35: 10236–10251.

36. Feldman D (2012) The Spike-Timing Dependence of Plasticity. Neuron 75: 556–571.

37. Kim T, Oh WC, Choi JH, Kwon HB (2016) Emergence of functional subnetworks in layer 2/3 cortex induced by sequential spikes in vivo. Proceedings of the National Academy of Sciences 113: E1372–E1381.

38. Ocker GK, Litwin-Kumar A, Doiron B (2015) Self-Organization of Microcircuits in Networks of Spiking Neurons with Plastic Synapses. PLoS Comput Biol 11: e1004458.

39. Trousdale J, Hu Y, Shea-Brown E, Josi K (2012) Impact of Network Structure and Cellular Response on Spike Time Correlations. PLoS Computational Biology 8: e1002408.

40. Kempter R, Gerstner W, Van Hemmen JL (1999) Hebbian learning and spiking neurons. Physical Review E 59: 4498.

41. Song S, Sjöström PJ, Reigl M, Nelson S, Chklovskii DB (2005) Highly nonrandom features of synaptic connectivity in local cortical circuits. PLoS Biol 3: e68.

42. Fourcaud-Trocme N, Hansel D, van Vreeswijk C, Brunel N (2003) How spike generation mechanisms determine the neuronal response to fluctuating inputs. Journal of Neuroscience 23: 11628–11640.

43. Markram H (1997) Regulation of Synaptic Efficacy by Coincidence of Postsynaptic APs and EPSPs. Science 275: 213–215.

44. Bi Gq, Poo Mm (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. The Journal of Neuroscience 18: 10464–10472.

45. Song S, Miller KD, Abbott LF (2000) Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. Nature Neuroscience 3: 919–926.

46. Van Rossum MC, Bi GQ, Turrigiano GG (2000) Stable hebbian learning from spike timing-dependent plasticity. The Journal of Neuroscience 20: 8812–8821.

47. Lubenov EV, Siapas AG (2008) Decoupling through synchrony in neuronal circuits with propagation delays. Neuron 58: 118–131.

48. Morrison A, Aertsen A, Diesmann M (2007) Spike-Timing-Dependent Plasticity in Balanced Random Networks. Neural Computation 19: 1437–1467.

49. Vogels TP, Froemke RC, Doyon N, Gilson M, Haas JS, et al. (2013) Inhibitory synaptic plasticity: spike timing-dependence and putative network function. Frontiers in Neural Circuits 7.

50. Vogels TP, Sprekeler H, Zenke F, Clopath C, Gerstner W (2011) Inhibitory Plasticity Balances Excitation and Inhibition in Sensory Pathways and Memory Networks. Science 334: 1569–1573.

51. Barron H, Vogels T, Emir U, Makin T, OShea J, et al. (2016) Unmasking latent inhibitory connections in human cortex to reveal dormant cortical memories. Neuron 90: 191–203.

52. Damour J, Froemke R (2015) Inhibitory and Excitatory Spike-Timing-Dependent Plasticity in the Auditory Cortex. Neuron 86: 514–528.

53. Ledoux E, Brunel N (2011) Dynamics of networks of excitatory and inhibitory neurons in response to time-dependent inputs. Frontiers in Computational Neuroscience 5: 25.

54. Babadi B, Abbott LF (2013) Pairwise Analysis Can Account for Network Structures Arising from Spike-Timing Dependent Plasticity. PLoS Computational Biology 9: e1002906.

55. Beck J, Bejjanki VR, Pouget A (2011) Insights from a simple expression for linear fisher information in a recurrently connected population of spiking neurons. Neural computation 23: 1484–1502.

56. Wei Y, Koulakov AA (2014) Long-term memory stabilized by noise-induced rehearsal. The Journal of Neuroscience 34: 15804–15815.

57. Bell CC, Han VZ, Sugawara Y, Grant K (1997) Synaptic plasticity in a cerebellum-like structure depends on temporal order. Nature 387: 278–281.

58. Rubin JE, Gerkin, RC, Bi, G-Q, Chow, C (2005) Calcium Time Course as a Signal for Spike-Timing-Dependent Plasticity. Journal of Neurophysiology 93: 2600–2613.

59. Shouval HZ, Wang SSH, Wittenberg GM (2010) Spike Timing Dependent Plasticity: A Consequence of More Fundamental Learning Rules. Frontiers in Computational Neuroscience 4.

60. van Vreeswijk C, Sompolinsky H (1998) Chaotic balanced state in a model of cortical circuits. Neural Computation 10: 1321–1371.

61. Renart A, de la Rocha J, Bartho P, Hollender L, Parga N, et al. (2010) The asynchronous state in cortical circuits. Science 327: 587–590.

62. Tetzlaff T, Helias M, Einevoll GT, Diesmann M (2012) Decorrelation of neural-network activity by inhibitory feedback. PLoS Comput Biol 8: e1002596.

63. Helias M, Tetzlaff T, Diesmann M (2014) The Correlation Structure of Local Neuronal Networks Intrinsically Results from Recurrent Dynamics. PLoS Comput Biol 10: e1003428.

64. Tsodyks MV, Skaggs WE, Sejnowski TJ, McNaughton BL (1997) Paradoxical effects of external modulation of inhibitory interneurons. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 17: 4382–4388.

65. Rubin DB, Van Hooser SD, Miller KD (2015) The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. Neuron 85: 402–417.

66. Miller KD, MacKay DJC (1994) The Role of Constraints in Hebbian Learning. Neural Computation 6: 100–126.

67. Miller KD (1996) Synaptic economics: competition and cooperation in synaptic plasticity. Neuron 17: 371–374.

68. Zenke F, Hennequin G, Gerstner W (2013) Synaptic Plasticity in Neural Networks Needs Homeostasis with a Fast Rate Detector. PLoS Comput Biol 9: e1003330.

69. Harnack D, Pelko M, Chaillet A, Chitour Y, van Rossum MC (2015) Stability of Neuronal Networks with Homeostatic Regulation. PLoS Comput Biol 11: e1004357.

70. Sweeney Y, Hellgren Kotaleski J, Hennig MH (2015) A Diffusive Homeostatic Signal Maintains Neural Heterogeneity and Responsiveness in Cortical Networks. PLoS Comput Biol 11: e1004389.

71. Zohary E, Shadlen MN, Newsome WT (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. Nature 370: 140–143.

72. Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. Nat Rev Neurosci 7: 358–66.

73. Kohn A, Coen-Cagli R, Kanitscheider I, Pouget A (2016) Correlations and neuronal population information. Annual review of neuroscience .

74. Abbott LF, Dayan P (1999) The Effect of Correlated Variability on the Accuracy of a Population Code. Neural Computation 11: 91–101.

75. Sompolinsky H, Yoon H, Kang K, Shamir M (2001) Population coding in neuronal systems with correlated noise. Physical Review E 64: 051904.

76. Josi K, Shea-Brown E, Doiron B, de la Rocha J (2009) Stimulus-dependent correlations and population codes. Neural Computation 21: 2774–2804.

77. Ecker AS, Berens P, Tolias AS, Bethge M (2011) The effect of noise correlations in populations of diversely tuned neurons. J Neurosci 31: 14272–14283.

78. Moreno-Bote R, Beck J, Kanitscheider I, Pitkow X, Latham P, et al. (2014) Information-limiting correlations. Nature Neuroscience 17: 1410–1417.

79. Hu Y, Zylberberg J, Shea-Brown E (2014) The Sign Rule and Beyond: Boundary Effects, Flexibility, and Noise Correlations in Neural Population Codes. PLoS Comput Biol 10: e1003469.

80. Zylberberg J, Cafaro J, Turner MH, Shea-Brown E, Rieke F (2016) Direction-selective circuits shape noise to ensure a precise population code. Neuron 89: 369–383.

81. Franke F, Fiscella M, Sevelev M, Roska B, Hierlemann A, et al. (2016) Structures of neural correlation and how they favor coding. Neuron 89: 409–422.

82. Aertsen A, Gerstein G, Habib M, Palm G (1989) Dynamics of neuronal firing correlation: modulation of "effective connectivity". J Neurophysiol 61: 900–917.

83. Ahissar E, Vaadia E, Ahissar M, Bergman H, Arieli A, et al. (1992) Dependence of cortical plasticity on correlated activity of single neurons and on behavioral context. Science (New York, NY) 257: 1412–1415.

84. Espinosa IE, Gerstein GL (1988) Cortical auditory neuron interactions during presentation of 3-tone sequences: effective connectivity. Brain Research 450: 39–50.

85. Ruff DA, Cohen MR (2014) Attention can either increase or decrease spike count correlations in visual cortex. Nature Neuroscience 17: 1591–1597.

86. Doiron B, Litwin-Kumar A, Rosenbaum R, Ocker GK, Josić K (2016) The mechanics of state-dependent neural correlations. Nature neuroscience 19: 383–393.

87. Kanitscheider I, Coen-Cagli R, Pouget A (2015) Origin of information-limiting noise correlations. Proceedings of the National Academy of Sciences 112: E6973–E6982.

88. Tsodyks M, Kenet T, Grinvald A, Arieli A (1999) Linking Spontaneous Activity of Single Cortical Neurons and the Underlying Functional Architecture. Science 286: 1943–1946.

89. Arieli A, Sterkin A, Grinvald A, Aertsen A (1996) Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. Science (New York, NY) 273: 1868–1871.

90. Kenet T, Bibitchkov D, Tsodyks M, Grinvald A, Arieli A (2003) Spontaneously emerging cortical representations of visual attributes. Nature 425: 954–956.

91. Luczak A, Barth P, Harris KD (2009) Spontaneous Events Outline the Realm of Possible Sensory Responses in Neocortical Populations. Neuron 62: 413–425.

92. Han F, Caporale N, Dan Y (2008) Reverberation of recent visual experience in spontaneous cortical waves. Neuron 60: 321–327.

93. Eagleman SL, Dragoi V (2012) Image sequence reactivation in awake V4 networks. Proceedings of the National Academy of Sciences 109: 19450–19455.

94. Xu S, Jiang W, Poo Mm, Dan Y (2012) Activity recall in a visual cortical ensemble. Nature Neuroscience 15: 449–455.

95. Graupner M, Brunel N (2007) STDP in a Bistable Synapse Model Based on CaMKII and Associated Signaling Pathways. PLoS Comput Biol 3: e221.

96. Higgins D, Graupner M, Brunel N (2014) Memory Maintenance in Synapses with Calcium-Based Plasticity in the Presence of Background Activity. PLoS Comput Biol 10: e1003834.

97. Seriès P, Latham PE, Pouget A (2004) Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. Nature neuroscience 7: 1129–1135.

98. Jolivet R, Lewis TJ, Gerstner W (2004) Generalized Integrate-and-Fire Models of Neuronal Activity Approximate Spike Trains of a Detailed Model to a High Degree of Accuracy. Journal of Neurophysiology 92: 959–976.

99. Jolivet R, Schrmann F, Berger TK, Naud R, Gerstner W, et al. (2008) The quantitative single-neuron modeling competition. Biological Cybernetics 99: 417–426.

100. Pernice V, Staude B, Cardanobile S, Rotter S (2011) How structure determines correlations in neuronal networks. PLoS Comput Biol 7: e1002059.

101. Cox D, Isham V (1980) Point Processes. Monographs on Statistics and Applied Probability. CRC Press.

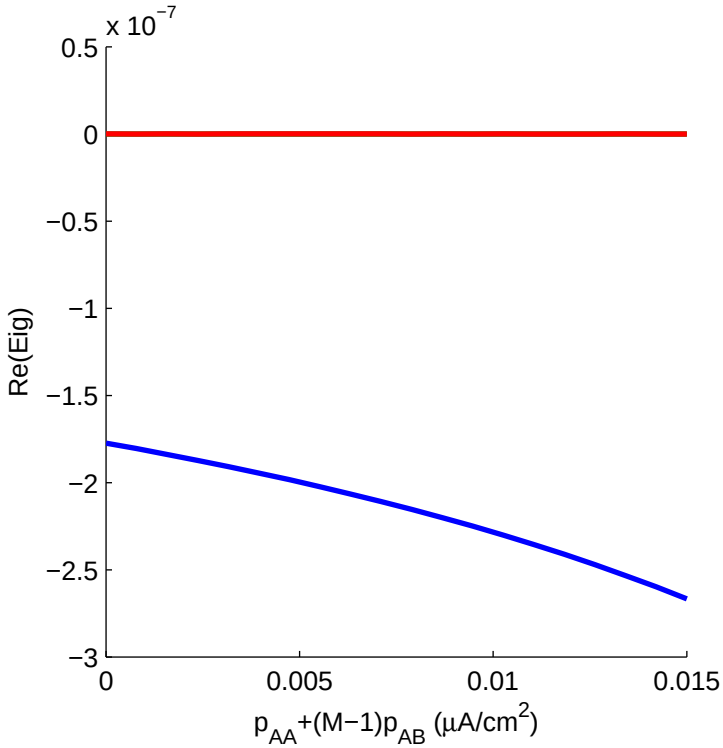

Figure 8. Supplemental Figure: Eigenvalues of the system Eqs. (4)-(50) **with unbalanced iSTDP.**